THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par *l'Université Toulouse III - Paul Sabatier*
Discipline ou spécialité : *Statistique Génétique*

Présentée et soutenue par *Mohamad Saad*
Le *5 Juillet 2012*

Titre : *Méthodes statistiques et stratégies d'études d'association de phénotypes complexes: Etudes pan-génomiques de la maladie de Parkinson*

JURY

Monsieur	*Pascal Sarda*	*Président*
Madame	*Florence Demenais*	*Rapporteur*
Monsieur	*Bertram Müller-Myhsok*	*Rapporteur*
Madame	*Françoise Clerget-Darpoux*	*Examinateur*
Madame	*Maria Martinez*	*Directrice de thèse*

Ecole doctorale : *Biologie Santé Biotechnologies*
Unité de recherche : *INSERM, UMR1043*
Directeur(s) de Thèse : *Maria Martinez*
Rapporteurs : *F. Demenais, B. Müller-Myhsok*

Mohamad SAAD

2012

Coordonnées du laboratoire d'accueil
INSERM U1043 - CPTP
CHU Purpan, BP 3028
31024 Toulouse CEDEX
France

E-mail : mohamad.saad@inserm.fr

mohamad.saad@live.fr

A ma mère …

Remerciements

Je tiens à exprimer mes remerciements les plus sincères dans un premier temps à ma directrice de thèse, Maria Martinez, pour m'avoir accueilli au sein de son équipe et pour m'avoir accordé toute sa confiance. Je la remercie également de m'avoir donné l'opportunité d'assister à plusieurs congrès internationaux et de participer à différents projets de recherche et de collaborations nationales et internationales. Je lui suis également reconnaissant pour ses qualités scientifiques et sa rigueur. Pendant ces trois années de formation, j'ai vécu une expérience enrichissante autant au niveau professionnel qu'humain, j'ai beaucoup appris à ses côtés et je lui adresse toute ma gratitude et admiration. Maria Martinez, sans votre aide, ce travail n'aurait jamais vu le jour.

Je tiens à exprimer mes remerciements les plus sincères aux rapporteurs de cette thèse, Florence Demenais et Bertram Müller-Myhsok, pour avoir accepté de rapporter ce travail de thèse, pour l'intérêt qu'ils y ont apporté et pour leurs conseils avisés.

Je remercie les membres du jury, Françoise Clerget-Darpoux et Pascal Sarda, pour l'attention qu'ils ont accordée à ce travail.

Je voudrais remercier également l'association France Parkinson pour avoir financé ma dernière année de thèse, ce qui m'a permis de terminer mon travail.

Je souhaite remercier vivement tous les membres de l'unité, pour leur sympathie et leur accueil spécialement Marie-Paule Roth, Hélène Coppin et François Canonne-Hergaux.

Je tiens à remercier nos collaborateurs, Alexis Brice, Suzanne lesage et Yannick Allanore.

Dans la vie de tous les jours, certaines personnes marquent leurs touches et gravent des souvenirs non-oubliables dans nos vies.

Toi, Nadid, tu en fais partie. J'ai appris plein de choses de toi. Ton comportement magnifique, tes réactions, juste parfaites, dans tous les moments, te rendent un être adorable et unique. J'admire ta curiosité scientifique et ton envie d'apprendre d'un domaine qui n'est pas le tien. Lorsque tes amis viennent me parler des gènes que nous avons identifiés, je ne peux que rester la bouche bée !! Je t'adresse un grand merci … Tu étais toujours là, lorsqu'il fallait. Tu m'as accompagné dans les moments difficiles de ma thèse. Là où je baissais les bras, tu me les remontais.

Je tiens à remercier ASP pour sa présence dans notre équipe, son aide et son soutien. I would like to thank NB pour son apport scientifique et aussi linguistique. Je n'oublie pas de remercier MM pour les moments agréables et les discussions variées qu'il apportait. Je respecte énormément ton respect, MM!

Un grand merci aux amis, proches et loin. Surtout ceux qui comptent beaucoup pour moi : HK, HS, HS, MC, SK, BN et OZ. Vous allez vous reconnaître surement !!

Pour terminer, je remercie ma famille pour son soutien permanent au cours de ces longues années d'études et d'absence, tout particulièrement mon père. Tu es mon guide dans cette vie et ma source d'inspiration. Père, je te suis reconnaissant pour tout ce que tu as fait. Sans toi, je ne serais jamais où je suis maintenant.

Résumé

Mon travail de thèse s'intéresse aux méthodes statistiques et stratégies d'étude de la composante génétique de maladies complexes chez l'homme et spécifiquement de la Maladie de Parkinson (MP). Ces travaux sont principalement développés dans le cadre d'études d'association pan-génomiques dans deux contextes : détection de variants fréquents et détection de variants rares. Le criblage du génome entier (GWAS) est une stratégie d'étude optimale à condition de bien contrôler les niveaux des erreurs de type I et de type II. En effet, un grand nombre de tests statistiques sont réalisés ; des problèmes de stratification de population sont possibles et leurs effets doivent être contrôlés. Par ailleurs, malgré leurs tailles d'échantillon relativement importantes, les études GWAS, basées sur le test simple-marqueur, peuvent s'avérer individuellement peu puissantes pour détecter des variants génétiques fréquents à effets faibles. L'utilisation des tests multi-marqueur peut optimiser l'utilisation de la variabilité génétique et donc augmenter la puissance des études GWAS. Je me suis intéressé à l'étude de ces tests et spécifiquement le test « SNP-Set » basé sur la méthode statistique de noyau et le test haplotypique. J'ai étudié les aspects théoriques de ces tests et j'ai évalué leurs propriétés statistiques dans nos données empiriques de la MP. Ainsi pour nos analyses de la MP, j'ai développé des techniques d'imputations et de méta-analyses afin d'augmenter la couverture de la variabilité génétique et la taille d'échantillon.

L'analyse d'association pour des variants rares présente plusieurs défis. Le test d'association simple-marqueur ne permet pas d'étudier tels variants et le coût des analyses à grande échelle de données de séquence reste prohibitif pour l'étude de maladies complexes. Notre design d'étude est une approche alternative qui repose sur la combinaison de données publiques de séquence aux données GWAS. Différents tests d'association pour l'étude de variants rares ont été récemment proposés mais leurs propriétés statistiques sont à ce jour mal connues. Par ailleurs, à l'échelle pan-génomique, les erreurs de type I et de type II de ces méthodes peuvent être influencées par certains facteurs comme la longueur du gène, l'hétérogénéité allélique dans le gène, le LD entre SNPs, le chevauchement entre gènes et la corrélation SNPs fréquents et maladie. J'ai évalué les propriétés statistiques de plusieurs de ces méthodes dans des données simulées et aussi dans nos données de la MP. Nous montrons que plusieurs méthodes, basées sur le modèle linéaire mixte, sont mathématiquement équivalentes et que certaines sont des cas particuliers d'autres. En conclusion, nous avons développé des stratégies et méthodes d'analyse, combinant des approches complémentaires (Maladie commune-variant fréquent vs Maladie commune –variant rare) dans le but d'optimiser la caractérisation de la composante génétique de la MP en particulier et de maladies complexes en générale.

Mots-clés : Statistique génétique, études d'association pan-génomiques, variants fréquents, variants rares, méthode de noyau, stratification de population, analyse en composante principale, imputation, haplotype, modèle linéaire, modèle linéaire mixte, régression multivariée.

Abstract

My thesis has focused on statistical methods and strategies to study the genetic components of complex human traits and especially of Parkinson's Disease (PD). My work was developed mainly in two contexts of genome wide association studies (GWAS): the detection of common variants and the detection of rare variants. GWAS is an optimal approach in which we have to control for the type I error and the type II error rates. Indeed, a large number of tests are performed. In addition, we must control for potential population stratification problems. Despite the large sample sizes in recent GWASs based on the single-marker test, they may have individually low power to detect common variants with small effects. The use of the multi-marker test may optimize the coverage of genetic variability and thus increase the power of GWAS. I have focused on the study of these tests, especially the "SNP-Set" test based on kernel machine regression and the haplotypic test. I studied the theoretical aspects of these tests and I evaluated the statistical properties in our empirical data for PD. In addition, in our analyses for PD, I developed imputation and meta-analysis techniques to increase the coverage of the genetic variability and the sample size.

Association analysis for rare variants faces several challenges. The single marker test is not powerful to detect such variants and the cost of whole-genome sequence analyses for complex traits is still prohibitive. Our design is a cost-effective alternative which is based on the joint use of public sequence data and GWAS data. Several new tests have been proposed but, to date, their statistical properties are still unclear. On the genome-wide level, the type I error and the type II error rates may depend on several factors as gene length, allelic heterogeneity in the gene, LD between SNPs, overlap between genes and the correlation between the common variants and the trait. I evaluated the statistical properties of several methods in simulated data and also in our GWAS PD data. We show that several methods, based on the linear mixed model, are mathematically equivalent and some are special cases of others. In conclusion, we developed strategies and analytical methods which combine complementary approaches (Common Disease-Common Variant versus Common Disease-Rare Variant) to optimize the characterization of the genetic components of PD in particular and of complex traits in general.

Keywords: Statistical genetics, genome wide association study, common variant, rare variant, kernel method, population stratification, principal component analysis, imputation, haplotype, linear model, linear mixed model, multivariate regression.

Production scientifique

Articles publiés

1) **Saad M**, Lesage S, Saint-Pierre A, Corvol JC, Zelenika D, Lambert JC, Vidailhet M, Mellick GD, Lohmann E, Durif F, Pollak P, Damier P, Tison F, Silburn PA, Tzourio C, Forlani S, Loriot MA, Giroud M, Helmer C, Portet F, Amouyel P, Lathrop M, Elbaz A, Durr A, Martinez M and Brice A, Genome-wide association study confirms BST1 and suggests a locus on 12q24 as the risk loci for Parkinson's disease in the European population. *Hum Mol Genet*, 2011. 20(3): p. 615-27

2) **Saad M**, Saint-Pierre A, Bohossian N, Macé M and Martinez M, Comparative study of statistical methods for detecting association with rare variants in exome-resequencing data. *BMC Proc*, 2011. 5(Suppl 9):S33

3) Allanore Y, **Saad M**, Dieudé P, Avouac J, Distler JH, Amouyel P, Matucci-Cerinic M, Riemekasten G, Airo P, Melchers I, Hachulla E, Cusi D, Wichmann HE, Wipff J, Lambert JC, Hunzelmann N, Tiev K, Caramaschi P, Diot E, Kowal-Bielecka O, Valentini G, Mouthon L, Czirják L, Damjanov N, Salvi E, Conti C, Müller M, Müller-Ladner U, Riccieri V, Ruiz B, Cracowski JL, Letenneur L, Dupuy AM, Meyer O, Kahan A, Munnich A, Boileau C and Martinez M, Genome-wide scan identifies TNIP1, PSORS1C1, and RHOB as novel risk loci for systemic sclerosis. *PLoS Genet*, 2011. 7(7): p. e1002091

4) Nalls MA, Plagnol V, Hernandez DG, Sharma M, Sheerin UM, **Saad M**, Simón-Sánchez J, Schulte C, Lesage S, Sveinbjörnsdóttir S, Stefánsson K, Martinez M, Hardy J, Heutink P, Brice A, Gasser T, Singleton AB, Wood NW, Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet*, 2011. 377(9766): p. 641-9

5) Plagnol V, Nalls MA, Hernandez DG, Sharma M, Sheerin UM, **Saad M**, Simón-Sánchez J, Schulte C, Lesage S, Sveinbjörnsdóttir S, Stefánsson K, Martinez M, Hardy J, Heutink P, Brice A, Gasser T, Singleton AB, Wood NW, A two-stage meta-analysis identifies several new loci for Parkinson's disease. *PLoS Genet*, 2011. 7(6): p. e1002142

6) Manetti M, Allanore Y, **Saad M**, Fatini C, Cohignac V, Guiducci S, Romano E, Airó P, Caramaschi P, Riccieri V, Bombardieri S, Abbate R, Caporali R, Cuomo G, Valesini G, Dieudé P, Hachulla E, Cracowski JL, Tiev K, Letenneur L, Chiocchia G, Ibba-Manneschi L, Martinez M, and Matucci-Cerinic M, Evidence for caveolin-1 (CAV1) as a new susceptibility gene regulating tissue fibrosis in systemic sclerosis. *Ann Rheum Dis*, 2012. doi:10.1136/annrheumdis-2011-200986

Communications orales

1) A Comparative Study of Statistical Methods for Detecting Association with Rare Variants in Exome-Resequencing Data. **Saad M.** Genetic Analysis Workshop 17, Octobre 2010, Boston, USA.
2) A genome-wide association study of Parkinson's disease. **Saad M.** Journée Régionale GenoToul Bioinfo, Mars 2011, Toulouse, France.

3) Statistical Methods for Detecting Association with Rare Variants in Exome-Resequencing Data. **Saad M.** Journée Régionale GenoToul Bioinfo, Mars 2011, Toulouse, France.
4) Rare Variants Association Study of common complex traits and diseases. **Saad M.** StatSeq workshop, Avril 2011, Toulouse, France.

Communications affichées

1) On the value of family data in genome-wide association studies for quantitative traits. **Saad M.** European Mathematical Genetic Meeting, Mai 2009, Munich, Germany.
2) Hunting for rare susceptibility variants in Genome Wide Association data on Parkinson's disease. **Saad M.** International Genetic Epidemiology Society, Septembre 2011, Heidelberg, Germany.
3) An evaluation of several statistical approaches to detect rare variants in Genome Wide Association data on Parkinson's disease. **Saad M.** 12th International Congress of Human Genetics, Octobbre 2011, Montréal, Canada.

ANNEXES

Annexe1: Genome-wide association study confirms BST1 and suggests a locus on 12q24 as the risk loci for Parkinson's disease in the European population.

Annexe 2: Comparative study of statistical methods for detecting association with rare variants in exome-resequencing data.

Annexe 3: Genome-wide scan identifies TNIP1, PSORS1C1, and RHOB as novel risk loci for systemic sclerosis.

Annexe 4: Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies.

Annexe 5: A two-stage meta-analysis identifies several new loci for Parkinson's disease.

Contexte et objectifs

On sait aujourd'hui qu'un grand nombre de pathologies a une composante génétique. Les maladies génétiques sont classiquement subdivisées en deux grandes catégories : maladies monogéniques et maladies multifactorielles.

Pour les maladies de type monogénique, comme par exemple, la maladie de Huntington, la mucoviscidose ou la phénylcétonurie, un seul facteur génétique est en cause. De plus, il est généralement le principal facteur de risque, l'influence de l'environnement étant souvent minime. Ainsi, la part des facteurs génétiques dans le déterminisme de la maladie, appelée héritabilité, est souvent très importante.

Les maladies multifactorielles sont causées conjointement par des facteurs génétiques et environnementaux et aussi très probablement par des effets d'interaction entre ces facteurs. Ces maladies se caractérisent par un âge de début de la maladie tardif et variable : elles concernent majoritairement le jeune adulte, voir la personne âgée. Elles sont nombreuses : diabète de type 2, maladies cardiovasculaires, cancers, maladie d'Alzheimer, maladie de Parkinson, sclérose en plaques, polyarthrite rhumatoïde, maladie de Crohn, schizophrénie, autisme, etc. Elles sont fréquentes dans la population et constituent donc un enjeu majeur de santé publique.

L'objectif de l'épidémiologie génétique est de caractériser la composante génétique des maladies chez l'homme, d'identifier les facteurs génétiques impliqués et d'estimer leurs effets propres et d'interaction. Elle repose sur le développement de méthodes statistiques qui intègrent, principalement, les marqueurs génétiques. En conséquence, cette discipline évolue en parallèle avec les progrès techniques en biologie. Le principe général est d'évaluer la corrélation entre le(s) marqueur(s) génétique(s) et le phénotype étudié (maladie ou trait quantitatif). Les marqueurs génétiques apportent deux types d'information : celle de liaison génétique, au niveau familial et celle d'association, au niveau de la population. L'unité d'échantillonnage des études de l'épidémio-génétique peut donc être de deux types : des sujets apparentés (paires de frères/sœurs, familles nucléaires, familles étendues/généalogies) ; sujets non apparentés (étude de type cas-témoins). Ces unités d'échantillonnage peuvent être aléatoirement sélectionnées ou non, par exemple au travers d'individus présentant le phénotype extrême (malades).

L'objectif des analyses de liaison génétique est de localiser les régions contenant les gènes responsables du trait ou de la maladie sur le génome dans des échantillons de familles. Sa

puissance dépend de la fréquence et de la pénétrance (effet) de l'allèle de susceptibilité, de la fréquence des phénocopies (malades non expliqués par le gène de susceptibilité) et des fréquences alléliques du(es) marqueur(s) étudié(s). La puissance est maximale lorsque la fréquence de l'allèle de susceptibilité est faible, la pénétrance est forte et le taux de phénocopies est faible (correspondance phénotype-génotype est peu ambigüe). La puissance dépend aussi de la taille des familles et de la distribution du phénotype au sein des familles. Ainsi, le design d'étude le plus performant est d'analyser des échantillons de familles sélectionnées via des individus de phénotype extrême. Dans le cas d'une maladie, il s'agira de recenser des familles contenant un grand nombre de malades et, si possible, sur plusieurs générations. Cette méthode a été largement utilisée pour l'étude des maladies monogéniques et a permis de cartographier un grand nombre des gènes impliqués dans ces pathologies. Cette méthode s'est également avérée puissante pour l'étude de certains sous-types de maladies multifactorielles, comme par exemple les formes à âge de début précoce de la maladie d'Alzheimer, de Parkinson ou du cancer du sein. L'analyse de liaison génétique s'avère être un outil puissant même lorsque plusieurs mutations au même gène sont impliquées (hétérogénéité allélique) : un bon exemple est celui du gène PSEN2 (presenilin) où plus d'une dizaine de mutations à transmission autosomique dominante expliquent une parties des formes précoces de la maladie d'Alzheimer. Notons que ce succès, dans l'étude de sous-type de certaines maladies multifactorielles, repose sur l'existence de grandes généalogies de malades, même si globalement elles ne représentent qu'une très faible minorité des cas de la population générale. Quoi qu'il en soit, que ce soit pour l'étude de maladies monogéniques ou de formes rares de maladies multifactorielles, il est clair que l'analyse de liaison parvient facilement à détecter la corrélation très forte qui existe entre le gène et la maladie. En revanche, elle est nettement moins adaptée pour l'étude de maladies multifactorielles où l'hétérogénéité génétique (plusieurs gènes impliqués) est très vraisemblable et où les effets des facteurs génétiques impliqués sont faibles (faible corrélation trait-marqueur). Par ailleurs, la prévalence relativement élevée de ces pathologies suggère que les allèles à risque sont fréquents dans la population générale.

L'analyse d'association en population est une alternative pertinente pour l'étude des maladies multifactorielles. Elle est fondée sur l'existence de dépendances statistiques, appelées déséquilibre de liaison, (linkage disequilibium), généralement observées entre les marqueurs proches (inférieures à 50-100 kilobases) sur l'ADN. Cette caractéristique biologique assure une localisation fine des mutations causales non observées à l'aide des marqueurs génétiques. Le principe est d'identifier les marqueurs génétiques pour lesquels les fréquences alléliques

différent significativement entre les malades et les témoins. La première vague de ces études portait sur des marqueurs de gènes candidats, gènes sélectionnés *à priori* selon des hypothèses biologiques. Ces études se sont avérées, le plus souvent, peu concluantes. Il existe quelques succès comme l'identification de l'association entre l'allèle e4 du gène APOE et le risque de la forme âgée de la maladie d'Alzheimer. Les avancées récentes des technologies génomiques à haut-débit ont permis d'accéder à une grande part de la variabilité du génome entier à l'aide de centaines de milliers de marqueurs. Ceux-ci sont des polymorphismes d'un seul nucléotide (SNP). Ces avancées technologiques et la mise en place du projet international « HapMap » ont ouvert la voie aux études d'association pan-génomiques, c'est-à-dire, la recherche d'association sur le génome entier (« Genome Wide Association Study », GWAS). Dans ce contexte, un certain nombre de grands projets de recherche ont vu le jour pour la caractérisation de la composante génétique de pathologies complexes. La première GWAS a été publiée en 2005. Elle a rapporté une association entre la dégénérescence maculaire liée à l'âge et un polymorphisme commun du gène codant pour le complément du facteur H. Par la suite, d'autres succès ont été rapportés avec notamment l'identification de variants impliqués dans différentes maladies complexes, comme le cancer de sein et le cancer de prostate. Néanmoins, la dimension et la complexité des données issues de ce nouveau type d'étude posent de nombreux défis statistiques.

L'approche GWAS est une stratégie d'étude optimale pour détecter les variants fréquents à condition de bien contrôler les niveaux des erreurs de type I et de type II. En effet, un grand nombre de tests statistiques sont réalisés ; des problèmes de stratification de population sont possibles et leurs effets doivent être contrôlés. Par ailleurs, malgré la taille des échantillons relativement importante, les études GWAS peuvent s'avérer individuellement peu puissantes. Pour augmenter la puissance, plusieurs méthodes existent comme l'imputation et la méta-analyse.

Il convient aussi de moduler notre enthousiasme. Ainsi, les variants identifiés à ce jour ont le plus souvent des effets faibles sur le risque de la maladie et n'expliquent qu'une faible part des cas. Il est clair que la variabilité du nucléotide n'est pas le seul type de variabilité du génome. Les variations structurelles comme les variations de nombre de copies (CNV) ou l'épi-génétique échappent aux études d'association pan-génomiques. Par ailleurs, des facteurs de l'environnement, seuls ou en interactions avec des facteurs génétiques, peuvent aussi expliquer une partie non négligeable du risque de la maladie. Nous nous limiterons ici à la caractérisation du risque expliqué par la variabilité ponctuelle de l'ADN. Dans ce contexte, il

est important de noter que les résultats GWAS sont plus probablement obtenus sous l'association indirecte que directe : les études GWAS identifient des signaux associations et non le(s) variant(s) causal(ux). Ces signaux reflètent la corrélation entre les allèles du SNP et ceux du variant causal, non observé. Ainsi, il est probable que la variabilité génétique commune testée n'est pas aussi complète que prévue. Ceci est certain en ce qui concerne la variabilité moins commune et, bien sur, rare. Idéalement, pour que l'association soit directe il faut observer le variant causal. Ceci requiert d'avoir la séquence entière du génome de plusieurs milliers de malades et témoins, ce qui reste impossible à cause du coût prohibitif de l'étude. Pour augmenter la probabilité de l'association directe, on peut, alternativement, améliorer la couverture de la variabilité génétique locale en utilisant conjointement plusieurs marqueurs génétiques. En effet, les études GWAS sont conduites avec des tests d'association simple-marqueur. Les tests d'association multi-marqueur classiques, comme le test multivarié ou le test haplotypique, souffrent de la relation croissante entre le nombre de marqueurs inclus et le nombre de degrés de liberté du test. A ce jour, une seule étude pan-génomique d'association multi-locus, basée sur le test haplotypique, a été publiée. D'autres tests multi-marqueur (« SNP-Set test») ont été récemment proposés. Ils présentent l'avantage de ne dépendre que d'un seul degré de liberté, comme le test classique simple-marqueur. Par ailleurs, comme nous le notons plus bas, ces tests sont également proposés pour la détection de l'association de variants rares. La question centrale est celle d'évaluer le gain potentiel apporté par l'information jointe de plusieurs marqueurs par rapport à l'analyse d'un seul marqueur.

L'hypothèse de variants peu fréquents, voir rares, à effets importants sur le risque des maladies multifactorielles soulève un certain engouement récent parmi la communauté scientifique. Les variants rares sont supposés avoir une origine plus récente que les variants communs et certains suggèrent qu'ils pourraient collectivement expliquer une part importante du risque de la maladie. Cette hypothèse est confortée par l'identification de gènes contenant des variants à risque rares comme dans l'exemple du gène IFIH1. Ce gène contient quatre variants rares (MAF < 3%) qui décroissent le risque de développer le diabète de type 1. L'analyse statistique d'association pour des variants rares pose des défis spécifiques plus importants que ceux des variants fréquents. Différents tests d'association ont été proposés. Leur approche générale est de combiner les allèles rares des variants d'une même unité génomique (le gène) en une seule variable. Les extensions ont été introduites pour filtrer ou non les allèles à combiner et/ou pour pondérer ou non les contributions individuelles des allèles rares du gène. Les tests multi-locus « SNP-Set » peuvent également être utilisés en

filtrant les allèles sur leurs fréquences. A ce jour, les propriétés statistiques de ces tests ont été évaluées principalement dans des données simulées. En dehors de la problématique statistique posée par l'étude d'association de variants rares, se pose le problème de l'accès à des données de séquençage à grande échelle pour des milliers de sujets. Ainsi, pour les maladies complexes, le coût du séquençage du génome ou de l'exome entier d'un grand nombre de sujets reste l'entrave essentielle à la recherche d'association de variants rares. Un design d'étude alternatif, appelé pseudo-séquençage, a été récemment proposé. Il repose sur la combinaison de données publiques de séquence aux données génotypiques de GWAS à travers des techniques d'imputation.

Les objectifs de cette thèse ont été d'évaluer différentes méthodes statistiques et design des études pan-génomiques. Les travaux se sont développés dans le cadre de l'analyse de données pan-génomiques de la maladie de Parkinson, dans le but d'améliorer notre compréhension de la composante génétique de cette pathologie en particulier, mais aussi d'autres maladies multifactorielles. Le manuscrit s'organise en trois grands chapitres.

Le premier chapitre constitue un rappel des notions et définitions de la génétique. La première partie décrit brièvement le matériel génétique, principe de la recombinaison, le déséquilibre de liaison et ses différentes mesures. La seconde partie introduit les études de l'épidémiologique génétique en décrivant les analyses de liaison paramétriques et non-paramétriques et les analyses d'association simple-marqueur et multi-marqueur.

Dans le deuxième chapitre, nous décrivons le contexte des études d'association pan-génomiques ainsi que leur robustesse et puissance pour la détection de la variation commune. Ensuite, nous décrivons l'épidémio-génétique de la MP et l'apport des analyses statistiques de liaison génétique et d'association pour cette pathologie.

Le troisième chapitre est composé de deux grandes sections dans lesquelles nous détaillons les travaux réalisés durant les trois années de thèse. Ces travaux sont principalement développés dans le cadre d'études d'association pan-génomiques dans deux contextes : recherche d'association de variations communes et de variations rares par l'approche simple-marqueur et par l'approche multi-marqueur, dans des données de génotypes, de pseudo-séquences et de séquences.

Dans la 1ère section, nous nous intéressons à la recherche d'association de variations communes dans des données de génotypes et de pseudo-séquences, par l'approche simple-marqueur. Nous exposons d'abord notre GWAS de la MP dans la cohorte Française (données

de génotypes). Nous décrivons le principe et les étapes des analyses de qualité-contrôle des données, les méthodes statistiques permettant de contrôler les effets de la stratification de population, les analyses d'association et, enfin, rapportons les principaux résultats.

Nous exposons ensuite notre étude de méta-analyse des données de l'IPDGC (données de pseudo-séquence). Nous décrivons le principe d'imputation de données (génotypes) manquantes et les méthodes statistiques de méta-analyse utilisées dans nos travaux.

Dans la 2ème section, nous nous sommes intéressés aux études pan-génomiques basées sur des tests multi-marqueur : le test haplotypique et le test « SNP-Set », basé sur le modèle du noyau. Nous décrivons différents modèles d'association multi-marqueur récemment proposés, et montrons les correspondances mathématiques entre certains de ces tests et « SNP-Set ». Nous avons cherché à savoir si le gain d'information permet de compenser l'augmentation du nombre de degrés de liberté du test haplotypique. Nous avons également cherché à quantifier le gain apporté par l'information jointe de plusieurs marqueurs vis-à-vis de l'analyse simple-marqueur. Ces évaluations ont été conduites sur les données pan-génomiques de notre étude française en utilisant les données de génotypes.

Nous exposons ensuite une étude dans laquelle nous avons évalué les propriétés statistiques de plusieurs tests d'association, spécifiquement développés pour l'étude de variants rares. Cette étude a été réalisée dans des données de séquençage, dans le cadre d'un atelier de travail sur les méthodes en génétique statistique « Genetic Analysis Workshop 17».

Table des matières

Chapitre 1

1. Définitions et généralités de la génétique et de l'épidémiologie génétique

1.1. Origine et structure de la variabilité du génome humain

- **Le matériel génétique**

L'homme est un organisme eucaryote pluricellulaire diploïde. En cela, le noyau de la majorité de ses cellules contient de l'ADN ou acide désoxyribonucléique en double dose. L'ADN est une macromolécule constituée de l'union de deux brins dits anti-sens ou antiparallèles ayant une structure spatiale en « double hélice ». Chaque brin est constitué de l'assemblage de constituants élémentaires appelés les nucléotides dont on dénombre quatre types: l'adénine (A), la guanine (G), la cytosine (C) et la thymine (T). Les deux brins s'associent entre eux au niveau de ces bases nucléotidiques en établissant des liaisons d'hydrogènes spécifiques. Ainsi, le nucléotide A d'un brin s'associe toujours avec T du brin complémentaire et C s'associe toujours avec G. L'ADN nucléaire de l'homme est composé de 3×10^9 paires de bases et constitue avec l'ADN mitochondrial ce que l'on appelle le génome humain. On estime que le génome humain contient entre 20,000 et 30,000 gènes. Un gène est défini comme une séquence d'ADN qui peut être transcrite en acide ribonucléique (ARN). Une autre définition est fondée sur l'idée qu'un gène correspond à une protéine à laquelle correspond une fonction précise. Les gènes représentent ~25 % du génome humain. Ils sont constitués de deux parties principales : les parties codantes transcrites en ARN appelé exons (~1% du génome humain) et les parties non-codantes et non transcrites en ARN appelées introns (~24% du génome humain). Le reste du génome contient des séquences non-codantes diverses comme les pseudo-gènes, les répétitions dispersées, les répétitions en tandem, etc [1].

- **Les marqueurs génétiques**

Les marqueurs génétiques sont des séquences d'ADN dont les positions exactes sur le génome sont connues. Ils représentent en quelque sorte des balises qui vont permettre de déterminer la localisation des loci responsables de la susceptibilité génétique à une maladie. Les premiers marqueurs génétiques furent les Restriction Fragment Length Polymorphisms (RFLPs), les Variable Number of Tandem Repeat (VNTRs ou minisatellites), les Short Tandem Repeat Polymorphisms (STRPs ou microsatellites) ou les Single Nucleotide

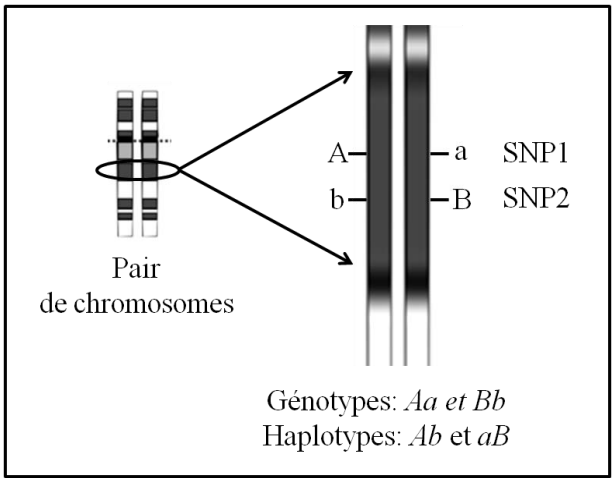
Polymorphisms (SNPs). Ils constituent une source d'information génétique riche et abondante. Les SNPs sont nombreux et apparaissant en moyenne une fois toutes les 100 à 300 bases. On estime que le génome humain contient entre 10 et 30 millions de SNPs. Dans la version d'Août 2010 du projet 1000Genomes, près de12 millions SNPs sont identifiés.

- Génotype et Haplotype

Le génotype est la composition allélique (nombre d'allèles) du marqueur, du locus ou du gène. Pour un locus bi-allélique, ayant deux allèles *A* et *a*, chaque individu est porteur d'un de trois génotypes possibles : *AA, aa* (homozygote) ou *Aa* (hétérozygote). L'haplotype est l'arrangement linéaire des allèles sur le même chromosome à deux (ou plus) locus (Figure 1.1).

Pour illustrer la différence entre haplotype – ou génotype phasé— et génotype, prenons l'exemple d'un individu homozygote au locus A (AA) et hétérozygote au locus B (Bb). Son génotype est AABb. Cette observation permet de déduire la phase allélique, i.e., elle permet d'établir la paire d'haplotypes: AB|Ab. En revanche, si l'individu est hétérozygote aux deux locus, l'observation de son génotype AaBb ne permet pas d'établir la distinction entre les deux combinaisons de paires d'haplotypes possibles : AB|ab et Ab|aB. Actuellement, seule la connaissance du génotype est accessible à faible coût. Pour connaitre la phase allélique, des méthodes statistiques d'inférence des haplotypes à partir des génotypes ont été développées [2,3,4].

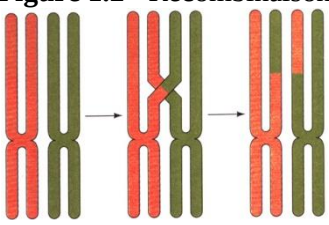
Figure 1.1- Notion de génotypes et d'haplotypes



- **La recombinaison génétique**

La recombinaison est un phénomène résultant du mélange de matériel génétique qui se produit par enjambement entre chromosomes. Elle survient au cours de la méiose, le processus de formation des gamètes mâles, les spermatozoïdes, et des gamètes femelles, les ovules. Chaque chromosome a alors la possibilité d'échanger une partie d'ADN avec son chromosome homologue (Figure 1.2). Le taux de recombinaison θ entre deux locus est la proportion des gamètes recombinés parmi l'ensemble des gamètes transmis par les parents. Il varie entre 0 et ½.

Figure 1.2- Recombinaison entre chromosomes homologues



On représente la probabilité de recombinaison entre deux locus par l'unité de distance génétique : centiMorgan (cM). 1 cM correspond à environ 1% de recombinaison c'est-à-dire une recombinaison en moyenne pour 100 méioses. L'équivalence entre la distance génétique et la distance physique entre deux locus varie selon l'espèce considérée. Chez l'homme, on admet : $1cM \sim 1Mb$ (Méga Base).

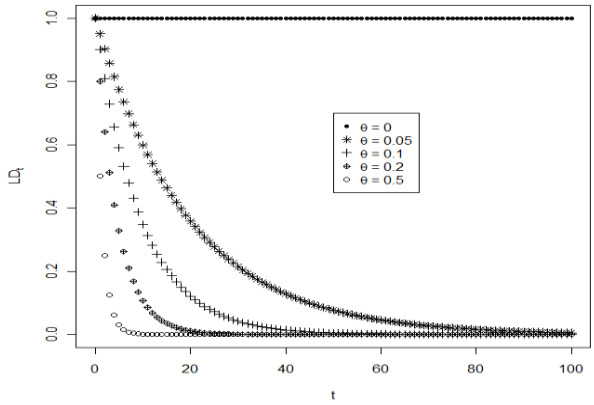
- **Le déséquilibre de liaison**

Le déséquilibre de liaison (Linkage Disequilibrium) est l'association non-aléatoire des allèles de deux (ou plus) locus au sein d'une population donnée. Ce terme a été proposé en 1960 par Lewontin et Kojim [5]. Le LD a beaucoup d'importance dans la biologie de l'évolution et dans la génétique humaine. Il apporte des informations sur les événements passés comme l'histoire de la population, le système de reproduction et le modèle de la division géographique. Dans une région chromosomique donnée, le LD peut refléter la sélection naturelle, la conversion de gènes, la mutation et d'autres forces qui causent l'évolution des fréquences alléliques. La relation de LD entre deux SNPs diminue avec le temps. Cette diminution dépend du taux de recombinaison locale θ :

$$LD_t = (1-\theta)^t \times LD_0$$

où LD_t est le LD au temps t et LD_0 est le LD initial (Figure 1.3).

Figure 1.3- Variation du LD avec le taux de recombinaison au cours du temps



Plus les deux locus sont proches, moins il y a de chance de recombinaison et plus le déséquilibre de liaison est maintenu au cours des générations. Le LD décroît donc avec la distance entre locus.

Mesures du déséquilibre de liaison pour 2 loci

Considérons une paire de SNPs ayant les allèles *A*/*a* et *B*/*b*. Notons f_A, $f_a = 1 - f_A$ et f_B, $f_b = 1 - f_B$ les fréquences alléliques correspondantes. Si un individu a le génotype *AA* au premier SNP et le génotype *Bb* au deuxième SNP, les deux haplotypes possibles sont *AB* et *Ab*. Notons par f_{AB} la fréquence de l'haplotype *AB*. Le calcul de LD entre l'allèle *A* et l'allèle *B* dépend des fréquences haplotypiques et alléliques comme le montre l'équation suivante:

$$D_{AB} = f_{AB} - f_A \times f_B$$

Le LD entre les autres allèles (Table 1.1) est calculé de la même façon. En faisant un simple réarrangement des allèles, on peut voir que $D_{AB} = D_{aB} = D_{Ab} = D_{ab} = D$.

Table1.1- Fréquences alléliques et haplotypiques

SNP1 \ SNP2	*B*	*b*	
A	f_{AB}	f_{Ab}	f_A
a	f_{aB}	f_{ab}	f_a
	f_B	f_b	1

La mesure *D* dépend fortement des fréquences alléliques. La comparaison du LD, et donc de la force de l'association, entre différentes paires de SNPs est donc difficile. Deux autres

mesures normalisées par rapport aux fréquences alléliques ont été introduites [6]. La première, *D'* [7], est définie comme suit :

$$|D'| = \begin{cases} \dfrac{-D_{AB}}{\min(f_A f_B, f_a f_b)} & si\ D_{AB} < 0 \\ \dfrac{D_{AB}}{\min(f_A f_b, f_a f_B)} & si\ D_{AB} > 0. \end{cases}$$

La deuxième mesure est le coefficient de corrélation de Pearson :

$$r^2 = \frac{D_{AB}^2}{f_A f_a f_B f_b}.$$

Lorsque $|D'| = 1$, on parle d'un déséquilibre de liaison complet. Cela se traduit par l'absence d'au moins d'un haplotype (Table 1.2).

Table 1.2- Fréquences alléliques et haplotypiques lorsque $|D'| = 1$

SNP1 \ SNP2	*B*	*b*	
A	f_{AB}	f_{Ab}	f_A
a	f_{aB}	0	$f_a = f_{aB}$
	f_B	$f_b = f_{Ab}$	1

D' = 1

Lorsque $r^2 = 1$, et donc forcément $|D'| = 1$, on parle d'un déséquilibre de liaison parfait. Cela se traduit par l'absence de deux haplotypes. Dans ce cas, les deux marqueurs apportent la même information (Table 1.3).

Table 1.3- Fréquences alléliques et haplotypiques lorsque $|D'| = 1\ et\ r^2 = 1$

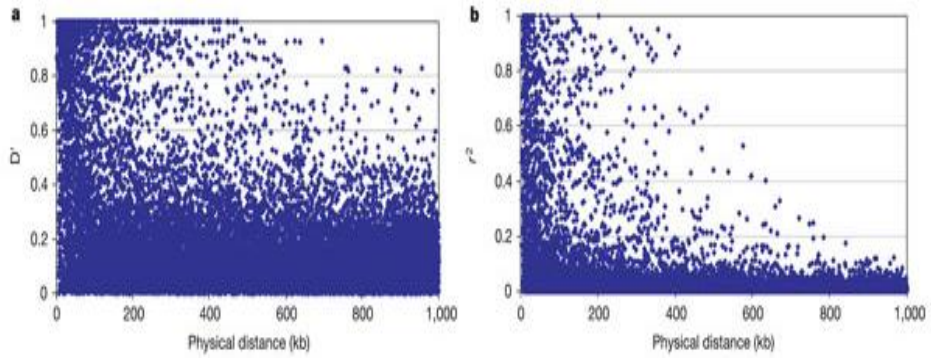
SNP1 \ SNP2	*B*	*b*	
A	0	f_{Ab}	$f_A = f_{Ab}$
a	f_{aB}	0	$f_a = f_{aB}$
	$f_B = f_{aB}$	$f_b = f_{Ab}$	1

D' = 1 et $r^2 = 1$

- **Les blocs de LD**

Les études récentes montrent une structuration du génome en fonction du LD résultant d'un non-homogène taux de recombinaison le long du génome. Ainsi, des groupes de locus sont transmis intacts de générations en générations. Ces groupes de marqueurs sont appelés bloc de LD ou bloc haplotypique. Ces blocs sont caractérisés par un fort degré de LD et une faible diversité haplotypique. En général, le LD ne s'étend pas au-delà de 50 kb. Cependant, des fortes valeurs de LD peuvent être observées au delà de 500 kb (Figure 1.4).

Figure 1.4 – La relation entre le LD et la distance physique sur le chromosome 22 : a) *D'* et b) r^2. Cette figure est tirée des travaux de *Dawson et al, Nature, 2002* [8].



Il y a des régions dans le génome dans le lequel le LD s'étend à plus de 5 Mb comme la région du complexe majeur d'histocompatibilité *CMH* (i.e. *HLA*) sur le chromosome 6. On appelle ce type de région: les régions de longs intervalles de LD. Plusieurs exemples sont cités dans [9].

Les tagSNPs – des marqueurs particuliers

Un tagSNP est un SNP corrélé à un ou plusieurs SNPs de son voisinage. Cette corrélation est mesurée par le paramètre r^2, vu précédemment. Comme nous l'avons vu, si cette corrélation est forte ($r^2 = 1$), l'information apportée par tous les SNPs du groupe est la même que celle apportée par un seul de ces SNPs. Ainsi, il n'est pas nécessaire de génotyper tous les SNPs pour capturer la variabilité génétique locale. Ceci a servi de base au développement des puces commerciales de SNPs. L'idée est de minimiser le nombre de SNPs à génotyper tout en maintenant un bon niveau de couverture de la variabilité. Généralement, on dit qu'un SNP tague un autre SNP lorsque $r^2 > 0.8$.

1.2. Méthodes statistiques de détection de facteurs génétiques de susceptibilité

Pour détecter des facteurs génétiques de maladies, il existe deux approches méthodologiques: les études de liaison et les études d'association. Ces approches sont des outils puissants et complémentaires pour caractériser la composante génétique des maladies. Généralement, on utilise dans un premier temps les études de liaison dont l'objectif est de localiser les régions chromosomiques où se trouvent un ou plusieurs gènes impliqués dans la maladie. Dans un deuxième temps, les études d'association pour préciser plus finement l'emplacement du gène.

1.2.1 Analyses de liaison génétique

Il s'agit d'une analyse statistique effectuée dans des données familiales. Elle permet de tester l'indépendance de transmission des allèles de marqueurs et la maladie dans les familles, et si ce n'est pas le cas, de localiser le gène de la maladie par rapport au marqueur.

Le paramètre qui mesure la liaison est le taux de recombinaison θ décrit dans le premier chapitre. Lorsque le taux de recombinaison est grand, c'est-à-dire lorsqu'il s'approche de 0.5, nous concluons qu'il n'y a pas de liaison génétique entre les deux locus. En présence d'une liaison génétique, le taux de recombinaison est inférieur à 0.5. Plus les locus sont proches, plus θ tend vers 0. Il existe deux approches pour tester la liaison : l'approche paramétrique classique et l'approche non-paramétrique.

1.2.1.1 Approche paramétrique : Le LOD Score

Ce test requiert de connaitre les paramètres du modèle génétique du trait étudié et son mode de transmission. Pour une maladie ces paramètres sont représentés par le vecteur $\alpha = (p, P_{G1}, P_{G2}, P_{G3})$: où p est la fréquence de l'allèle à risque et P_{G1}, P_{G2}, P_{G3} , les trois pénétrances (probabilité d'être atteint conditionnellement au génotype)

$$\begin{cases} P_{G1} = \Pr(\text{malade}|aa)\,; \\ P_{G2} = \Pr(\text{malade}|Aa)\,; \\ P_{G3} = \Pr(\text{malade}|AA)\,. \end{cases}$$

Le test est basé sur le calcul de la vraisemblance de la liaison sachant les observations et le modèle génétique. La vraisemblance varie pour le paramètre θ, le taux de recombinaison entre le marqueur et le locus du trait sous-jacent. L'hypothèse de liaison est testée par le LOD score qui est le log en base dix du rapport des vraisemblances de l'hypothèse alternative (liaison génétique : $\theta = \theta_1 \leq 0.5$) et l'hypothèse nulle (pas de liaison: $\theta = 0.5$).

Soit *Y* le vecteur de phénotypes des *N* individus de la famille *F* à *n* enfants et *M* le marqueur étudié. La vraisemblance de θ est :

$$L(\theta|F) \propto P(F|\theta) = P(Y, M|\theta, \alpha) = P\left(Y_f, M_f, Y_m, M_m, Y_1, M_1, \dots, Y_n, M_n \middle| \theta, \alpha\right)$$
$$= P\left(Y_f, M_f\right)P(Y_m, M_m)P\left(Y_1, M_1, \dots, Y_n, M_n \middle| Y_f, M_f, Y_m, M_m, \theta, \alpha\right)$$
$$= \sum_{k=1}^{3} P\left(Y_f|G_{f,k}\right)P\left(G_{f,k}\right)P\left(M_f\right) \times \sum_{k=1}^{3} P\left(Y_m|G_{m,k}\right)P\left(G_{m,k}\right)P(M_m) \times$$
$$\prod_{o=1}^{n}\sum_{k=1}^{3} P\left(Y_o \middle| G_{o,k}, \alpha\right) \times P(G_{o,k}, M_o|G_{f,k}, M_f, G_{m,k}, M_m, \theta)$$

avec *f*=père et *m*=mère.

Dans la famille *i*, le test de LOD score s'écrit comme suit :

$$Z_i(\theta_1) = \log_{10}\left(\frac{L_i(\theta = \theta_1)}{L_i(\theta = 0)}\right).$$

Le LOD score d'un échantillon de *k* familles est la somme des LOD scores des *k* familles :

$$Z(\theta_1) = \sum_{i=1}^{k} Z_i(\theta_1).$$

Critère de décision

La procédure de test est de type séquentiel. On accumule de l'information (des familles) jusqu'au moment où il sera possible de trancher entre les hypothèses H_0 et H_1. La valeur du LOD score indique les probabilités relatives d'observer l'échantillon sous H_1 et sous H_0. Ainsi, un LOD score de 3 signifie que la probabilité d'observer l'échantillon est 1000 fois plus grande sous H_1 que sous H_0. Les critères de décisions sont :

- Si $Z(\theta_1) > 3$, on rejette H_0 et on conclut par la liaison ($\theta = \theta_1$) ;
- Si $Z(\theta_1) \leq -2$, on ne rejette pas H_0 et on conclut par l'absence de liaison ($\theta = 0$) ;
- Si $-2 < Z(\theta_1) \leq 3$, on ne peut pas trancher et il faut continuer à accumuler l'information.

Dans le cadre d'analyse de maladies mendéliennes, les paramètres du modèle peuvent être facilement spécifiés à partir des analyses de ségrégations. Si ces paramètres sont mal spécifiés, il en résulte une perte de puissance et aussi une augmentation d'erreur de type I. Or, nous ne savons pas spécifier a priori ces paramètres dans le cas des maladies complexes. Les analyses de liaison non-paramétriques ont donc été développées pour se libérer de ces contraintes.

1.2.1.2 Approche non-paramétrique

Le principe de ces analyses est de chercher s'il existe une corrélation entre la ressemblance au trait et la ressemblance au marqueur entre apparentés. La similarité génétique de deux individus est mesurée par ***l'identité par descendance (IBD)***. Deux allèles sont identiques par descendance s'ils sont copies d'un même allèle présent chez un ancêtre commun. La variable IBD mesure le nombre d'allèle IBD chez deux individus (IBD = 0, 1 ou 2).

Une des méthodes non-paramétriques proposées est celle de Haseman et Elston [10] basée sur des paires de germains. Pour un trait quantitatif, le modèle est une régression linéaire classique sous la forme : $\Delta_i = \left(y_{i,1} - y_{i,2}\right)^2 = \alpha + \beta\pi_i + \epsilon_i$,

avec $y_{i,1\,\&\,2}$ et π_i sont les phénotypes et la proportion d'allèle IBD (0,1 ou 2) chez les germains de la famille *i*. L'hypothèse nulle de non liaison est $\beta = 0$.

Pour un trait binaire, la méthode non-paramétrique est classiquement basée sur l'analyse de paires de germains atteints. Elle consiste à comparer la distribution observée de la proportion d'allèles IBD dans l'échantillon à la distribution attendue chez des germains (¼, ½, ¼ pour 0, 1, 2 allèles IBD). Si la liaison existe, on s'attend à un excès des paires de germains où l'IBD vaut 2. Ici, le test statistique peut être est un test de chi-2 de conformité.

1.2.2 Analyses d'association dans des données de population

Contrairement aux analyses de liaison, les analyses d'association peuvent être conduites dans des données de sujets apparentés ou non. L'approche populaire est celle des études en population (sujets non apparentés) et pour une maladie, celle de type cas-témoins. Les études d'association cherchent à mettre en évidence une différence dans les fréquences allèliques du marqueur entre les malades et les témoins.

1.2.2.1 Tests d'association simple-marqueur

Soit un SNP ayant les allèles *A* et *a*.

La table 1.4 montre les deux tableaux de contingence génotypique et allélique.

Table 1.4- Tableaux de contingence génotypique et allélique.

A. Table génotypique

	AA	*Aa*	*aa*	
Malades	*n11*	*n12*	*n13*	$n1. = N_{cas}$
Témoins	*n21*	*n22*	*n23*	$n2. = N_{tem}$
	$n.1=N_{AA}$	$n.2=N_{Aa}$	$n.3=N_{aa}$	$n = N$

B. Table Allélique

	A	*a*	
Malades	$m11 = 2\times n11+n12$	$m12 = 2\times n13+n12$	$m1. = 2N_{cas}$
Témoins	$m21 = 2\times n21+n22$	$m22 = 2\times n23+n22$	$m2. = 2N_{tem}$
	$m.1 = 2N_{AA} + N_{Aa}$	$m.2 = 2N_{aa} + N_{Aa}$	$m = 2N$

Odds Ratio

L'Odds Ratio (OR) est le rapport de deux risques relatifs : le risque d'observer l'allèle *A* chez les cas relativement aux témoins sur le risque d'observer l'autre allèle chez les cas relativement aux témoins. L'OR permet de mesurer la force de l'association entre la maladie et le SNP. Il est égal à:

$$OR = \frac{m_{11} \times m_{22}}{m_{21} \times m_{12}}.$$

L'odds ratio, comme toute estimation réalisée sur un échantillon, doit être présenté avec son intervalle de confiance à 95 %, qui mesure la précision de l'estimation. Cet intervalle peut être estimé en utilisant certaines méthodes comme celle de Woolf et de Miettinen. La méthode de Woolf consiste à estimer la variance du log(OR) qui suit une loi normal :

$$var(\log(OR)) = \frac{1}{m_{11}} + \frac{1}{m_{12}} + \frac{1}{m_{21}} + \frac{1}{m_{22}}.$$

L'intervalle de confiance à 95% est calculé comme suit :

$$IC95\% = \exp\left[\log(OR) \pm 1.96 \times \sqrt{var(\log(OR))}\right]$$

La méthode de Miettinen est une méthode simple qui permet de calculer l'intervalle de confiance de l'OR à partir des résultats du test de χ^2 d'association entre la maladie et le SNP:

$$IC95\% = \exp\left[\log(OR) \pm 1.96 \times \sqrt{\chi^2}\right]$$

- Test de χ^2 d'indépendance : Test allélique

Le test est basé sur la table de contingence allélique (table 1.4 (B)).

L'hypothèse *H0* se formule par : $\left\{ fij = \frac{mij}{m} = \frac{mi.}{m} \times \frac{m.j}{m} = fi.\times f.j \text{ avec } i = 1,2 \text{ et } j = 1,2 \right\}$

Le test correspondant est fondé sur la statistique de Pearson :

$$X_A = \sum_{i=1}^{2} \left(\frac{\left(m1i - \frac{m1.\times m.i}{m}\right)^2}{\frac{m1.\times m.i}{m}} + \frac{\left(m2i - \frac{m2.\times m.i}{m}\right)^2}{\frac{m2.\times m.i}{m}} \right)$$

$$\sim \chi^2(1 \text{ degré de liberté } df)$$

- **Test du χ^2 d'indépendance : Test génotypique**

Le test est basé sur la table de contingence génotypique (table 1.4 (A)).

L'hypothèse *H0* se formule par : $\left\{ pij = \frac{nij}{n} = \frac{ni.}{n} \times \frac{n.j}{n} = pi. \times p.j \text{ avec } i = 1,2 \text{ et } j = 1,2,3 \right\}$

Le test correspondant est fondé sur la statistique de Pearson :

$$X_G = \sum_{i=1}^{3} \left(\frac{\left(n1i - \frac{n1. \times n.i}{n}\right)^2}{\frac{n1. \times n.i}{n}} + \frac{\left(n2i - \frac{n2. \times n.i}{n}\right)^2}{\frac{n2. \times n.i}{n}} \right) \sim \chi^2_{(2)}$$

- **Test du rapport de vraisemblance (LRT)**

Dans le cas du test allélique, le test s'écrit comme suit :

$$LRT_A = -2 \sum_{i=1}^{2} \sum_{j=1}^{2} mij \log \left(\frac{mij}{\frac{mi. \times m.j}{m}} \right) \sim \chi^2_{(1)}$$

Dans le cas du test génotypique, le test s'écrit comme suit :

$$LRT_G = -2 \sum_{i=1}^{2} \sum_{j=1}^{3} nij \log \left(\frac{nij}{\frac{ni. \times n.j}{n}} \right) \sim \chi^2_{(2)}$$

- **Test Exact de Fisher**

Pour que le test de χ^2 soit valide, il faut avoir au moins cinq observations dans chaque cellule de la table de contingence. Sinon, il faut utiliser le test exact de Fisher. La probabilité d'observer la table 1.4.B est issue de la distribution hypergéométrique et est égale:

$$P(m11) = \frac{\binom{m1.}{m11}\binom{m2.}{m21}}{\binom{m}{m.1}}$$
$$= \frac{m1.! \times m2.! \times m.1! \times m.2!}{m11! \times m12! \times m21! \times m22! \times m!}$$

Le test consiste à calculer toutes les tables possibles en fixant les marginaux ($m1., m2., m.1$ et $m.2$), et en variant $m11$ de la façon suivante :

$$\max(0, m1. + m.1 - m) \leq m11 \leq \min(m1., m.1).$$

Les probabilités d'observer les tables possibles s'écrivent :

$$P(m11 = t) = \frac{\binom{m1.}{t}\binom{m2.}{m.1-t}}{\binom{m}{m.1}}.$$

Dans la table de contingence 2×2, l'indépendance de deux groupes correspond à un OR=1. L'hypothèse alternative, H1, peut être unilatérale $OR > 1$ ou bilatérale $OR \neq 1$. La valeur-p du test unilatéral est :

$$valeur - P = \sum_{t \geq m11} P(t);$$

La valeur-p du test bilatéral est:

$$valeur - P = \sum_{P(t) \leq P(m11)} P(t).$$

- **Test de tendance : Cochran-Armitage [11,12]**

Le test est basé sur la table de contingence génotypique (Table 1.4 (A)).

L'hypothèse *H0* se formule par : $\{ p11 = p12 = p13 \}$. Le test statistique de tendance est :

$$X_T = \frac{n[n(n12 + 2n13) - n1.(n.1 + 2n.2)]^2}{n1.n2.[n(n.1 + 4n.2) - (n.1 + 2n.2)^2]} \sim \chi^2_{(1)}.$$

- **Régression Logistique**

Une autre façon de tester l'association est d'utiliser le modèle de régression logistique:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \alpha + X\beta \quad \text{(M1)}$$

avec $p = P(Y = 1|X = x) = 1 - P(Y = 0|X = x)$ et *Y=1* pour les cas et *0* pour les témoins.

Sous l'hypothèse que les observations sont indépendantes et pour un échantillon assez large, les estimations des paramètres ($\hat{\alpha}, \hat{\beta}$) se font par la méthode de maximum de vraisemblance.

Test d'hypothèse pour la régression logistique : En 1943, Wald a montré que les estimateurs de maximum de vraisemblance suivent une loi normale. Plusieurs tests d'hypothèse ont été donc proposés comme le test de Wald et le test de rapport de vraisemblance.

Pour tester l'hypothèse *H0* : $\beta = 0$, le test statistique de Wald est: $\mathrm{Z} = \frac{\hat{\beta}}{\sqrt{\mathrm{Var}(\hat{\beta})}} \sim N(0,1)$ sous *H0.*

Il faut noter que Z^2 suit asymptotiquement une loi de $\chi^2_{(1)}$.

Le test de rapport de vraisemblance s'écrit comme suit:

$$Z = -2\left[\log\big(L(Y, \hat{\alpha}, \beta = 0)\big) - \log\Big(L\big(Y, \hat{\alpha}, \hat{\beta}\big)\Big)\right]$$

Z suit asymptotiquement une loi de $\chi^2_{(1)}$ à 1 degré de liberté.

Asymptotiquement, les deux tests donnent les mêmes résultats. Cependant, le test de LRT est préférable au test de Wald [13] parce qu'il utilise plus d'information en incorporant le log de vraisemblance estimé sous *H0* ($\hat{\beta} = 0$) et sous *H1* ($\hat{\beta} \neq 0$). Pour des grandes valeurs de $|\hat{\beta}|$, le test de Wald est moins puissant que le LRT et peut donner des résultats aberrants.

- **Test de Maentel-Haenzel (MHT)**

Ce test est souvent utilisé lorsque les sujets analysés parviennent d'origines géographiques différentes. Le test se base sur les tables de contingence allélique de chaque strate (origine géographique) et calcule un OR globale. Soit h=1,…,*H* les différentes strates. Reprenons la table de contingence allélique (table 1.4 (B)) mais en ajoutant l'indice *h* à toutes les valeurs.

	A	*a*	
Malades	m_{h11}	m_{h12}	$m_{h1.}$
Témoins	m_{h21}	m_{h22}	$m_{h2.}$
	$m_{h.1}$	$m_{h.2}$	m_{h}

La statistique de test s'écrit comme suit :

$$\chi^2_{MHT} = \frac{(\sum_{h=1}^{H} m_{h11} - \sum_{h=1}^{H} k_{h11})^2}{\sum_{h=1}^{H} V_{h11}} \sim \chi^2_{(H-1)}$$

avec $V_{h11} = \frac{m_{h1.} \times m_{h2.} \times m_{h.1} \times m_{h.2}}{m_h^3 (m_h - 1)}$ et $k_{h11} = \frac{m_{h1.} \times m_{h.1}}{m_h}$.

L'OR globale est calculé comme suit :

$$OR_{MHT} = \frac{\sum_h \frac{m_{h11} \times m_{h22}}{m_h}}{\sum_h \frac{m_{h12} \times m_{h21}}{m_h}}.$$

Pour tester l'homogénéité des Ors entre les chaque strates, on peut utiliser le test de Breslow-Day. La statistique du test est :

$$\chi^2 = \sum_i \sum_j \sum_h \frac{\left(m_{hij} - k_{hij}\right)^2}{k_{hij}} \sim \chi^2_{(H-1)}$$

L'hypothèse nulle est : $H_0: OR_1 = \cdots = OR_h = \cdots = OR_H$.

1.2.2.2 Tests d'association multi-marqueur

- **Régression multivariée**

Prenons le modèle (M1) de la régression logistique simple marqueur. Dans la régression mutlivariée, la variable *X* est une matrice contenant les *p* SNPs à tester. Dans ce cas, β est un vecteur de taille *p* : $\beta = (\beta_1, \dots, \beta_p)$. L'hypothèse nulle d'absence d'association est $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$. Le test de Wald ou le LRT peuvent être utilisés. Le nombre de *df* est *p-1*. A cause de ce nombre qui augmente avec le nombre *p* de SNPs testés, le test de la régression multivariée peut manquer de puissance.

- **Test d'association haplotypique**

Lorsque les phases allèliques sont connues, on peut construire une table de contingence pour les fréquences des haplotypes chez les malades et chez les témoins (Table 1.5).

Table 1.5- Table de contingence haplotypique.

	H_1	H_2	…	H_k
Malades	n_{11}	n_{12}	…	n_{1k}
Témoins	n_{21}	n_{22}	…	n_{2k}

Le test haplotypique est, comme dans le modèle multivarié, un test de χ^2 à *k-1* df.

Cependant, comme déjà noté plus haut, en pratique seuls les génotypes sont observés et pas les haplotypes. L'inférence des phases alléliques dépend des génotypes observés, et donc des fréquences alléliques des SNPs. Différentes méthodes d'inférence des phases alléliques de données génotypiques ont été développées [2,3,4]. Les tests haplotypiques basés sur les données inférées sont possibles, comme celui implémenté dans PLINK [14], basé sur la régression linéaire. Cette approche tient compte de l'incertitude dans l'estimation des haplotypes. Pour un trait binaire, le même modèle logistique (M1) est utilisé. La matrice *X* s'écrit comme suit :

$$X = \begin{array}{c} \begin{matrix} H_1 & H_2 & \dots & H_k \end{matrix} \\ \begin{pmatrix} p_{i1} & p_{i2} & \dots & p_{ik} \end{pmatrix} \end{array}$$

avec $p_{ij} =$ Prob (haplotype *j* chez l'individu *i*)

1.2.3. Remarques et conclusions

- **Modèles génétiques**

Le modèle génétique général sous le modèle de régression est :

$$\text{logit}(p) = \alpha + \beta_1 x_1 + \beta_2 x_2 \text{ ;}$$
$$\text{H0 : } \hat{\beta}_1 = \hat{\beta}_2 = 0;$$

Le génotype *AA* sera codé par $\begin{pmatrix} x_1 = 0 \\ x_2 = 0 \end{pmatrix}$, *Aa* par $\begin{pmatrix} x_1 = 0 \\ x_2 = 1 \end{pmatrix}$, et *aa* par $\begin{pmatrix} x_1 = 1 \\ x_2 = 1 \end{pmatrix}$.
Le test d'association a 2 df.

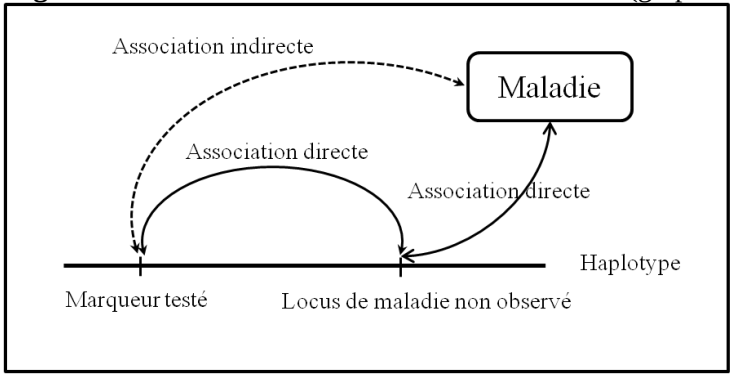
Alternativement, plusieurs sous-modèles peuvent être supposés: -additif sur l'échelle logarithmique du risque ($OR_{AA} = OR_{Aa}^2$) -dominant ($OR_{AA} = OR_{Aa}$) -récessif ($OR_{Aa} = OR_{aa}$. La puissance statistique du test d'association dépend du vrai modèle génétique, généralement inconnu pour les maladies complexes.

L'avantage de la régression logistique par rapport au test de χ^2 est 1) la possibilité d'inclure des covariables et 2) la facilité de tester les modèles génétiques emboités. Dans le modèle $\text{logit}(p) = \alpha + \beta x$, les génotypes *AA*/*Aa*/*aa* sont codés par (*x = 0*, *x = 1*, *x = 2*) sous le modèle additif, par (*x = 0*, *x = 1*, *x = 1*) sous le modèle dominant et par (*x = 0*, *x = 0*, *x = 1*) sous le modèle récessif, en supposant que *a* est l'allèle à risque. Le test de χ^2 peut être adapté au modèle dominant et récessif en regroupant les colonnes de la table de contingence. Le test de tendance de Cochran-Armitage suppose le modèle additif.

- **Association directe et indirecte :**

Généralement, les analyses d'association sont conduites dans le cadre d'association indirecte : le marqueur testé ne contribue pas à la variabilité du phénotype mais il se trouve physiquement proche du variant de susceptibilité et leurs allèles sont en LD.

Figure 1.5- Notion d'association directe et indirecte (graphe inspiré de David J. Balding [15]).



- **Équilibre d'Hardy Weinberg (HWE)**

Les tests d'association en population supposent que les distributions génotypiques du(es) marqueur(s) suivent la distribution sous l'équilibre de Hardy-Weinberg. L'équilibre de Hardy Weinberg repose sur trois hypothèses : (1) la population est de taille infinie, (2) il n'y a pas de migration, ni mutation, ni sélection, et enfin (3) il y a panmixie (les unions se font au hasard). L'absence d'un de ces facteurs peut entraîner une déviation par rapport à cet équilibre. D'autres facteurs peuvent créer une telle déviation comme des artefacts techniques et les erreurs de génotypage (discuté dans le Chapitre 3, section 3.1.1).

Pour tester l'écart à l'équilibre de Hardy Weinberg (HW), on peut utiliser le test du χ^2 de Pearson d'homogénéité entre les distributions génotypiques observées et celles attendues sous l'équilibre.

Soit *A* et *a* les deux allèles à un locus donné. Soit $\hat{f}_{AA}, \hat{f}_{Aa}, \hat{f}_{aa}$ les fréquences génotypiques estimées dans un échantillon de *N* sujets. Les fréquences (et les effectifs) génotypiques attendues, sous l'hypothèse nulle d'équilibre de HW, sont déduites des fréquences alléliques:

$$H_0 : \begin{cases} & \text{Fréquences} & \text{Effectifs} \\ f_{AA} = & f_A^2 & (f_A^2 \times N) \\ f_{Aa} = & 2f_A f_a & (2f_A f_a \times N) \\ f_{aa} = & f_a^2 & (f_a^2 \times N) \end{cases}$$

Pour tester la déviation par rapport à HWE, on peut utiliser la statistique suivante :

$$\chi^2 = \sum \frac{(\text{Effectif}_{Observé} - \text{Effectif}_{Attendu})^2}{\text{Effectif}_{Attendu}} \sim \chi^2_{(1)}$$

$$= \left[\frac{\left(N \times \hat{f}_{AA} - N \times f_A^2\right)^2}{N \times f_A^2} + \frac{\left(N \times \hat{f}_{Aa} - 2N \times f_A f_a\right)^2}{2N \times f_A f_a} + \frac{\left(N \times \hat{f}_{aa} - N \times f_a^2\right)^2}{N \times f_a^2} \right].$$

Sensibilité des tests d'association en population

Un avantage du design de l'étude en population est la facilité d'assembler un grand nombre de cas et de témoins indépendants. En revanche, sous ce design, les tests d'association sont particulièrement sensibles aux problèmes de mésappariement des cas aux témoins, aux variables de confusion et en particulier au problème d'une structure génétique cachée de la population [16,17]. Ce problème est évident lorsque les cas et les témoins proviennent de populations génétiquement différentes. Dans ce cas, de nombreux signaux d'association peuvent résulter à cause de la structure génétique différentielle. Pour éviter ce type de

problème, il convient donc de sélectionner les cas et les témoins à partir d'une même population et évidemment de bien les apparier sur les variables épidémiologiques associées à la maladie, comme le sexe et/ou l'âge pour les maladies de l'adulte. Cependant, comme discuté dans le chapitre suivant, une stratification résiduelle peut toujours exister et il faut en contrôler les effets. Le problème de stratification est évité dans les designs d'études familiales de type « case-parent trios » et « case-sibling » [18]. En plus, les membres de chaque famille sont exposés aux mêmes facteurs environnementaux qui peuvent être associés à la maladie. Le principal inconvénient de ces designs est la difficulté d'assembler de grandes familles surtout lorsque la maladie est peu familiale. A cela s'ajoute, la difficulté d'assembler de l'ADN des « case-parent » trios, pour une maladie de l'adulte.

Chapitre 2.

Etudes d'association pan-génomiques de la maladie de Parkinson

2.1 Contexte des études d'association pan-génomiques

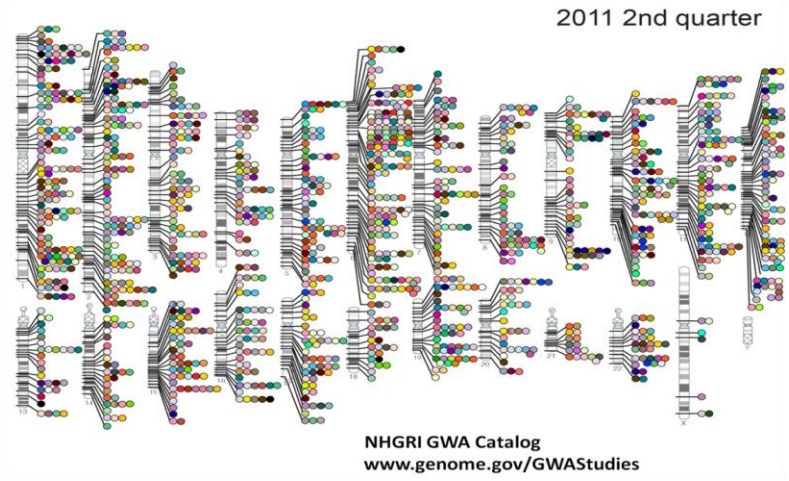
L'ère des études d'association a pris de l'importance sur les analyses de liaison à la fin des années 1990, avec la publication de Risch et Merikangas [19]. Ces auteurs ont suggéré que l'avenir des études génétiques des traits complexes serait l'analyse d'association systématique tout au long du génome. Sous l'hypothèse « ***Maladie Commune-Variant Commun*** » («Common Disease–Common Variant», CDCV), ce type d'étude était annoncé comme la solution de la cartographie génétique des traits complexes [19]. L'avancement des technologies de génotypage et la mise en place du projet HapMap [20] a permis d'effectuer les études d'association pan-génomiques. Le projet HapMap avait pour objectif de créer une carte d'un million de SNPs communs et rendre publique leurs fréquences alléliques, leurs génotypes ainsi que la relation de LD entre eux, dans trois populations (90 individus YRI = Yoruba, Ibadan-Nigeria ; 90 individus JPT= Japonais, Tokyo-Japon et CHB= Chinois, Pékin-Chine ; 90 individus CEPH/CEU= résidents aux états unis, originaires de l'Europe du nord et de l'Europe de l'ouest). Par la suite, le projet HapMap a été étendu pour inclure, dans la phase 2, plus de trois millions de SNPs génotypés chez les mêmes individus. La connaissance des blocs de LD le long du génome et des relations de LD entre les SNPs a permis de développer des cartes denses de tagSNPs. Le développement technologique a aussi rendu possible le génotypage à haut-débit. La combinaison de ces facteurs a permis le développement des études pan-génomiques à grande échelle à l'aide de puces commerciales de SNPs. La table 2.1 montre les cartes de tagSNPs de différentes puces commerciales : dans la population européenne, la couverture génomique de la puce Human1M est 93%. Cela signifie que près de 93% des SNPs communs du génome sont représentés (i.e., en fort LD ; $r^2 \geq 0.8$) par un ou plusieurs SNPs de la puce.

Table 2.1- Nombre de SNP des puces commerciales et pourcentage de la variabilité commune ($r^2 > 0.8$) qui est capturée par ces puces dans différentes populations de HapMap. Ces résultats ont été tirés de Li et al, [21].

SNP chips	# de SNPs	CEU	CHB+JPT	YRI
HumanHap300	317,511	77%	66%	29%
HumanHap550	555,352	87%	83%	50%
HumanHap650Y	660,917	87%	84%	60%
Human1M	1,072,820	93%	92%	68%
SNP Array 5.0	500,568	64%	66%	41%
SNP Array 6.0	934,968	83%	84%	62%

La première étude d'association pan-génomique a été publiée en 2005. Elle a rapporté une association entre la dégénérescence maculaire liée à l'âge et un SNP commun du gène codant pour le complément du facteur H [22]. A ce jour, des centaines d'études d'association pan-génomiques ont été réalisées dans plus de 80 maladies et traits. Les études publiées sont répertoriées dans un catalogue en ligne mis à jour en continu : http://www.genome.gov/gwastudies/ (Figure 2.1).

Figure 2.1- Catalogue des locus associés avec des traits complexes



Robustesse et puissance des études d'association pan-génomiques

Dans les études d'association pan-génomiques, un grand nombre de test est réalisé. Le problème du test multiple est à considérer sérieusement pour décider d'un niveau adéquat de signification pour rejeter l'hypothèse nulle de non association. Des critères de significations plus stricts doivent donc être appliqués.

La figure 2.2 montre le nombre de sujets (N) qu'il faut étudier, pour avoir une puissance de 80% au seuil nominal de 10^{-6}, en fonction de la fréquence de l'allèle mineur (***MAF***) et de l'OR. Pour une MAF donnée, N augmente lorsqu'OR décroit; de même, pour un OR donné,

N augmente lorsque MAF décroit. En d'autres termes, la puissance est meilleure pour des grandes valeurs de l'OR (effet de l'allèle à risque) et/ou de la MAF (fréquence de l'allèle mineur).

Figure 2.2- Taille d'échantillon nécessaire pour détecter l'association entre la maladie et le variant causal avec une puissance de 80% au seuil $\alpha=10^{-6}$ en fonction de la taille de l'effet (OR= 1.2 - (rouge), 1.3 (bleu), 1.5 (jaune) et 2 (noir)) et de la MAF du variant causal. Figure tirée des travaux de Wang et al, [17].

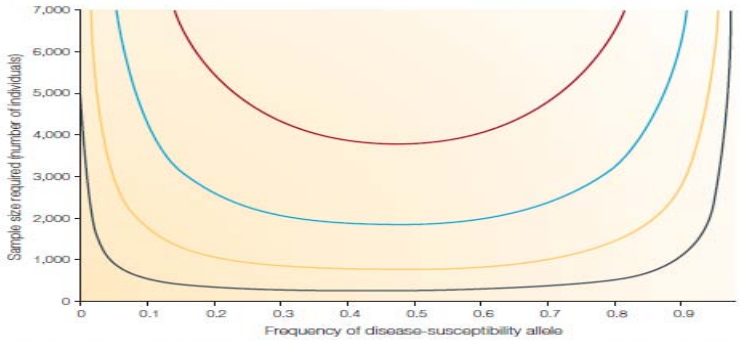


Figure 1 | **Effects of allele frequency on sample-size requirements.** The numbers of cases and controls that are required in an association study to detect disease variants with allelic odds ratios of 1.2 (red), 1.3 (blue), 1.5 (yellow) and 2 (black) are shown. Numbers shown are for a statistical power of 80% at a significance level of $P < 10^{-6}$, assuming a multiplicative model for the effects of alleles and perfect correlative linkage disequilibrium between alleles of test markers and disease variants.

Le niveau de signification pan-génomique défini par le Welcome Trust Consortium WTCCC est de 5×10^{-8}. Cela revient à supposer que la variabilité génétique peut être résumée par un million de SNPs indépendants (5×10^{-8} = 0.05/1000000). Pour obtenir une bonne puissance (>80%) à ce niveau de signification, des grand échantillons sont nécessaires (Figure 2.2). Lorsque la taille de l'échantillon et le nombre de tests sont très grands, un simple biais peut causer une augmentation importante du taux de faux positifs. Les études de réplication sont donc nécessaires. Une approche commune est celle des études GWASs en plusieurs étapes. Dans une première étape, on réalise le criblage du génome dans un échantillon relativement grand. Puis on retient les meilleurs signaux d'association pour étude de réplication dans un ou plusieurs échantillons indépendants qui doivent être issus de préférence de la même population que celle du premier échantillon.

2.2 Epidémio-génétique de la maladie de Parkinson

La maladie de Parkinson (MP) est la seconde maladie neurodégénérative la plus fréquente après la maladie d'Alzheimer. Elle se caractérise par une perte massive des neurones dopaminergiques dans la substance noire du cerveau [23]. Les principaux syndromes parkinsoniens sont le tremblement de repos, l'akinésie, la rigidité et l'instabilité postural [24].

La maladie est déclarée chez une personne si elle présente au moins deux syndromes parkinsoniens [25]. La plupart des patients reçoivent un diagnostique clinique de type MP possible ou MP probable [26]. La maladie de Parkinson touche environ 1% de la population des pays industrialisés âgée de plus de 65 ans. La prévalence de la maladie augmente avec l'âge et atteint le 4% dans la huitième et neuvième décennie de la vie [27]. L'âge moyen d'apparition de la maladie est d'environ 60 ans, même si pour 5 à 10% des cas, classés comme précoces, la maladie commence entre 20 et 50 ans [28]. Les hommes seraient plus souvent atteints que les femmes [29].

L'étiologie de la MP est complexe et reste mal connue. Cependant, plusieurs études ont rapporté une association avec des facteurs environnementaux comme l'exposition aux pesticides, herbicides et aux métaux lourds [30,31]. La majorité des cas sont isolés. Les formes familiales représentent de 5 à 15% des cas. Des analyses de ségrégation ont montré que le risque de récurrence familiale λ_r varie entre 2 et 3 [32,33]. Il existe aussi des formes de la MP, encore plus rares, de familles contenant plusieurs atteints sur plusieurs générations. Ces formes sont généralement à âge de début précoce. Les analyses de liaison de LOD score, ont permis l'identification de dix locus de susceptibilité (PARK1-PARK10) et ensuite de plusieurs gènes de la maladie : alpha-synuclein (SNCA), parkin, ubiquitin-C-terminal hydrolase L1, Leucine-rich repeat kinase-2 (LRRK2), DJ-1, PINK1 et UCHL-1. On distingue les deux gènes SNCA et LRRK2 avec un mode de transmission autosomique dominant et parkin, PINK1, DJ-1 et ATP13A2 avec un mode de transmission autosomique récessif [34]. Par analyse d'association de type gène candidat, l'association de la MP avec deux gènes (MAPT et GBA) a été suggérée. Le gène MAPT est impliqué dans la maladie de dégénérescence lobaire fronto-temporale qui présente des syndromes parkinsoniens [35] et le gène GBA est impliqué dans la maladie de Gaucher [36]. En conclusion, la contribution génétique au risque de la MP reste majoritairement inconnue.

Les deux premières études GWAS de la MP [18,37] ont identifié plusieurs signaux d'association mais pas au niveau de signification pan-génomique. Les résultats n'ont pas été non plus confirmés [38]. Dans ces deux études, les tailles d'échantillons étaient faibles (i.e. 332/332 et 267/270 cas/témoins respectivement) ainsi que la densité des cartes de SNPs (198345 et 408000 respectivement). Au moment où nous avons conçu notre étude d'association pan-génomique dans les données Françaises, deux études GWAS [39,40] ont été publiées en 2009. Elles étaient basées sur une étape de criblage suivi d'une étape de réplication, chacune dans des grands échantillons. La densité des cartes de SNPs était elle

aussi améliorée par rapport aux 2 études pilotes. La première étude a été réalisée dans un échantillon de sujets issus des populations Nord-américaine, Anglaise et Allemande ; formant un échantillon global de 5074 cas et 8551 témoins [40]. La deuxième étude était réalisée chez des patients et témoins (N global : 2011 cas et 18381 témoins) issus de la population japonaise [39]. Cette étude a identifié l'association à quatre locus, au niveau de signification pan-génomique : SNCA (OR=1.37, P=7.35×10^{-17}), LRRK2 (OR=1.39, P = 2.72×10^{-8}), BST1 (OR = 1.24, P = 3.94×10^{-9}) et PARK16 (OR =1.3, P = 1.52×10^{-12}). La première étude n'a identifié que deux locus au niveau de signification pan-génomique : SNCA (OR =1.23, P = 2.24×10^{-16}) et MAPT (OR =0.77, P = 1.95×10^{-16}). Dans cette étude, les signaux d'association de plusieurs variants des gènes LRRK2 et PARK16 étaient statistiquement répliqués. En revanche, aucune association n'était rapportée avec les variants du gène BST1.

Chapitre 3

Tests d'association et designs d'études pan-génomiques

L'analyse d'association est fondée sur l'existence de dépendances statistiques, le LD, entre les marqueurs/variants de l'ADN. Ainsi, la probabilité de détecter l'association dépend fortement de la corrélation entre le variant testé et le variant causal. Elle est maximale si le variant testé est associé et a les mêmes fréquences alléliques que le variant causal (r^2=1) et nulle en absence de corrélation (r^2=0), même si le variant testé est situé au voisinage proche du variant causal. Lorsque la corrélation est complète (r^2=1), l'association est directe, sinon elle est dite indirecte.

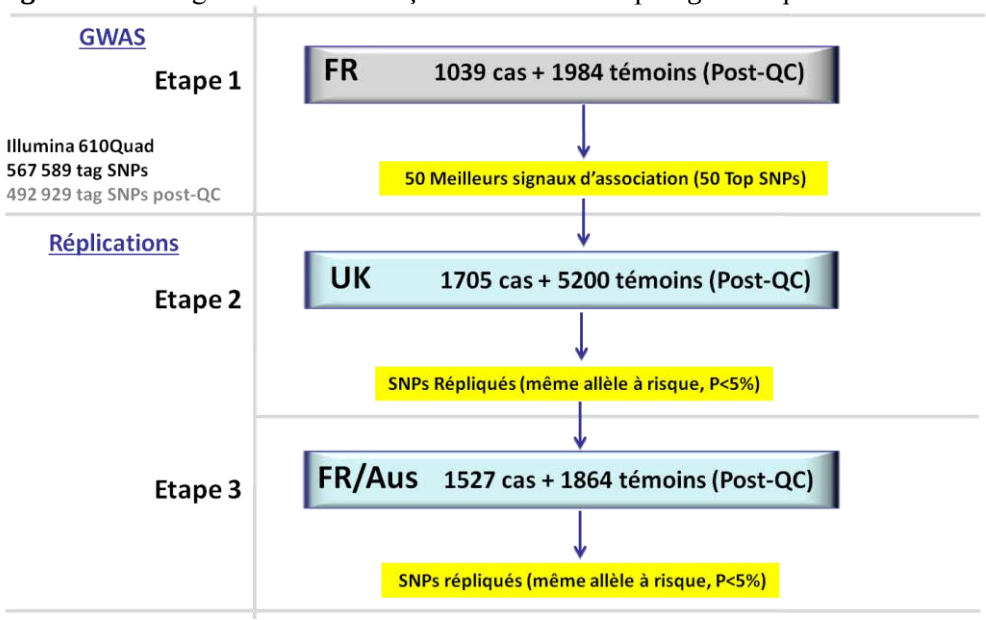
Le design classique des études pan-génomiques a été développé principalement pour la recherche d'association sous l'hypothèse « Maladie Commune – Variant Fréquent » (CD-CV). Ces études sont classiquement basées sur le test d'association simple-marqueur : la puissance du test pour détecter de l'association directe ou quasi-directe (r^2~0.8) de variants communs est bonne pour des tailles d'échantillon de plusieurs milliers de sujets (cf. Figure 2.2). En effet, comme nous l'avons vu dans le chapitre précédent, les puces commerciales de SNPs, utilisées pour les études de génotypage à haut-débit, permettent de capturer la majorité de la variabilité génétique commune du génome humain. Les SNPs de la puce sont des tagSNPs communs (MAF>0.1), fortement corrélés (r^2>0.8) avec un ou plusieurs autres SNPs/variants communs du génome mais absents de la puce. On reconnait, cependant, que pour certaines régions du génome, la couverture des puces de SNPs n'est pas aussi parfaite. Par ailleurs, ces puces ne couvrent pas ou peu la variabilité ponctuelle rare ou moins fréquente. Etant donné les grandes tailles d'échantillon (plusieurs dizaines de milliers de sujets) qui sont requises pour obtenir un bon niveau de puissance, au niveau de signification pan-génomique, il est clair que la recherche d'association pour des variants rares ou peu fréquents, n'est pas l'objectif premier des études GWAS.

3.1 Test simple-marqueur

3.1.1. Analyse GWAS française

Pour identifier de nouveaux variants à risque de la MP, nous avons conçu une étude d'association pan-génomique en trois étapes (N total, >13300 sujets). Le design est schématisé dans la figure 3.1. Les 50 signaux d'association identifiés dans la première étape (1039 cas et 1984 témoins français) ont été testés dans la deuxième étape, de réplication *in silico*, dans un échantillon de cas-témoins anglais du WTCCC (1705 cas et 5200 témoins) [41]. Les signaux positivement répliqués dans cette étape ont été de nouveau testés dans l'étape 3, de réplication *de novo*, dans une cohorte indépendante de cas-témoins français et australiens (1527 cas et 1864 témoins).

Figure 3.1- Design de l'étude Française d'association pan-génomique de la MP



- **Matériel et méthodes**

a/ Echantillons

Trois échantillons de malades et témoins ont été inclus dans notre étude.

Etape 1 – scan

1- Les patients (N=1064) ont été recrutés via le réseau français d'études génétiques de la maladie de Parkinson qui constitue 15 hôpitaux universitaires français. Un effort d'enrichissement de patients ayant une histoire familiale positive a été fait pour augmenter le nombre de patients ayant une prédisposition génétique à la MP.

2- Les témoins utilisés (N=2023) sont sélectionnés de la cohorte française de trois cités (3C : Bordeaux, Montpellier et Dijon) qui est une étude prospective en population. Cette cohorte contient plus de 9000 témoins issus de la population européenne [42].

Un descriptif détaillé des échantillons des patients et des témoins est montré dans la table suivante :

	Etape 1- scan	
	Patients	**Témoins**
N	1039	1984
Sexe: M/F	1.42	1.33
Age: Moyenne + SD (n)	57.5 ± 16.6 (1003)	73.7 ± 5.4 (1984)
AOO: Moyenne + SD (n)	48.9 ± 12.8 (970)	-
FH+ (%)	47	-

FH+: histoire familiale positive; M: masculin, F: féminin
AOO: âge de début de la maladie; SD: déviation standard

Etape 2 – réplication *in silico*

Dans cette étape, nous avons échangé les données pan-génomiques avec nos collaborateurs anglais [41] du WTCCC2 : 1705 cas et 5200 témoins issus de la cohorte « the 1958 Birth Cohort » et « the OK Blood Services Controls » [43].

Etape 3 – réplication *de novo*

Cette étape comprend deux échantillons indépendants. L'échantillon français (872 cas et 1440 témoins) provenait des cohortes TERRE et PARTAGE constituées par Alexis Elbaz [31]. Le deuxième échantillon a été constitué via nos collaborations, à partir d'une collection de malades et témoins issus de la population australienne (655 cas et 424 témoins).

b/ Génotypages

1- Dans l'étape de scan, les malades et les témoins ont été génotypés par le Centre National de Génotypage (CNG) avec la puce Illumina Human610-Quad BeadChip (~600000 SNPs).
2- Dans l'étude anglaise (réplication *in silico*), les sujets ont été génotypés avec la puce Illumina 650Y, par le Welcome Trust Center.
3- Les génotypages de cette étape (réplication *de novo*) ont été réalisés par des techniques de TaqMan, dans le laboratoire de Alexis Brice et Suzanne Lesage (Pitié-Salpetrière)

c/ Analyses de contrôle qualité (QC)

Les études d'associations pan-génomiques sont susceptibles de contenir des erreurs résultants d'artefacts techniques. Le génotypage à haut-débit requiert certaines conditions quant à la qualité d'ADN qui, si elles ne sont pas satisfaites peuvent conduire à des erreurs de

génotypage. Par ailleurs, des erreurs peuvent aussi être commises lors du transfert du tube d'ADN, de l'assignation du tube au sujet de l'étude, etc. Les analyses de qualité contrôle des données pan-génomiques ont pour but de minimiser l'impact de ces erreurs et artéfacts sur les analyses d'association. Cette étape se déroule en deux parties : le nettoyage des données des SNPs et celle des individus. Nous avons utilisé les critères classiquement proposés pour ces analyses.

En générale, on exclut les SNPs qui présentent un taux de succès (« Call Rate », CR) inférieur à 95%. La valeur seuil du CR peut dépendre de la fréquence de l'allèle mineur (MAF) du SNP : par exemple, les SNPs communs (MAF>0.05) sont exclus si le CR est < 95% ; les SNPs moins fréquents (MAF<0.05) si le CR est < 99%, car l'analyse d'association est plus sensible aux erreurs de génotypages. D'autres études proposent d'exclure tous les SNPs avec un MAF <1%, puisque la puissance de détecter l'association est, dans ce cas, faible. Une différence entre les cas et les témoins pour CR est un autre facteur important à contrôler. Il est recommandé d'exclure un SNP lorsque cette différence de CR entre les cas et les témoins est significative [44]. Souvent, cette différence apparaît lorsque les échantillons de cas et de témoins n'ont pas été génotypés ensemble. Dans notre étude, nous avons exclu un SNP si le MAF est < 5% ou le CR ≤ 97%. L'écart à l'équilibre de Hardy Weinberg des distributions génotypiques des SNPs dans l'échantillon est aussi testé. Toutefois, il est clair qu'une association vraie entre le trait et le marqueur peut créer une déviation par rapport à l'équilibre de Hardy Weinberg. Ainsi, la significativité à l'écart de HW est principalement évaluée dans l'échantillon des témoins. Les critères de signification généralement utilisés dans la littérature sont assez stricts pour tenir compte du nombre de tests réalisés. Dans notre étude, nous avons exclu un SNP si la valeur-P du test de HWE est $\leq 10^{-5}$ dans l'échantillon des témoins. Au total, cette première étape de contrôle qualité a conduit à l'exclusion de 74660 SNPs. Toutes les analyses ultérieures ont été basées sur 492929 SNPs (appelé SNPs postQC).

Lorsque l'ADN est de mauvaise qualité ou de faible concentration, son amplification peut échouer ou le signal fluorescent peut ne pas être assez fort pour déterminer le génotype. Il en résulte donc un grand nombre de données manquantes (génotypes manquants) associées à ce tube d'ADN (sujet). Il est important d'exclure les individus dont le taux de succès (CR) est faible car ceci peut suggérer que leurs génotypes sont assez imprécis. Des critères stricts d'exclusion d'individu, comme CR< 95% ou 99%, sont donc utilisés. Dans notre étude, 22 individus ont été exclus sur la base d'un CR < 96%. Les tests d'association en données de

population supposent que les sujets sont non apparentés. L'identification de doublons ou sujets apparentés peut se faire par différentes approches basées sur l'estimation de l'identité par descendance entre les sujets [14,45,46,47,48]. Par exemple, PLINK [14] implémente une approche de type chaîne de Markov cachée pour estimer les probabilités suivantes : $Z_0 = P(\text{IBD} = 0)$, $Z_1 = P(\text{IBD} = 1)$ et $Z_2 = P(\text{IBD} = 2)$. La proportion des allèles partagés par descendance entre deux individus sera donc :

$$\hat{\pi} = \frac{2 \times Z_2 + 1 \times Z_1 + 0 \times Z_0}{2} = Z_2 + 0.5 \times Z_1$$

Cette valeur varie entre 0 et 1. La valeur 1 signifie que l'individu est dupliqué (ADN envoyé deux fois) ou que les deux sujets sont des vrais jumeaux. Les valeurs de $\hat{\pi} = \frac{1}{2}, \frac{1}{4}$ et $\frac{1}{8}$ représentent un lien de parenté du premier, deuxième et troisième degré, respectivement. En général, le critère utilisé dans les études GWAS est 0.13=1/8. Dans notre étude, 39 sujets (14 cas et 25 témoins) ont été exclus en suivant le critère $\hat{\pi} > 0.13$.

d/ Analyses préliminaires– Recherche d'une stratification de population dans l'échantillon

Un facteur de confusion important est celui d'un possible mélange et/ou de stratification de population des cas et des témoins. Nous donnons, ici, une définition de la population au sens génétique telle que donnée par *Yuri S. Aulchenko* [49]: « *Deux individus I_1 et I_2 appartiennent à la même population si 1) la probabilité qu'ils aient un descendant en commun est plus grande que zéro et 2) cette probabilité est plus grande que la probabilité que I_1 et I_2 aient un descendant commun avec I_3 qui appartient à une autre population* ».

Certaines régions du génome sont constituées de SNPs, dont les fréquences alléliques sont population-spécifiques. Ainsi, la présence d'une stratification génétique dans l'échantillon de cas-témoins, entraine une augmentation du taux de fausses associations [16]. Dans les études GWAS, même avec un bon design d'étude et une bonne sélection des sujets, une stratification résiduelle peut toujours exister et peut donc impacter le test d'association [50,51].

Une première approche de correction, appelée le « Contrôle Génomique » (GC), a été proposée en 1999 par Devlin, B. Roeder, K. [52]. Cette correction repose sur le fait qu'on admet que seule une infime proportion des SNPs testés peut être impliquée dans la variabilité du trait. Dans ce cas, la médiane de la distribution de $\chi^2_{(1)}$ observée dans les données ne doit

pas trop s'écarter de la valeur attendue sous l'hypothèse nulle d'absence d'association. La valeur de la médiane théorique est 0.456. Pour évaluer l'existence de biais, on inspecte la distribution pan-génomique du test d'association, $\chi^2_{observé}$, et on identifie sa médiane. Le rapport de la médiane observée à la valeur théorique donne le coefficient du contrôle génomique, $\lambda = \frac{médiane(\chi^2_{observé})}{0.456}$.

Le coefficient λ indique le degré d'inflation de la distribution statistique et doit être proche de 1. Si ce n'est pas le cas, les valeurs observées du test d'association ($\chi^2_{observé}$) sont ensuite divisées par λ. Ce coefficient dépend de la taille d'échantillon et peut être normalisé pour un échantillon de 1000 cas et 1000 témoins. Le coefficient normalisé s'appelle $\lambda_{1000} = 1 + (\lambda_{obs} - 1) \times (\frac{1}{n_{cas}} + \frac{1}{n_{témoins}})/(\frac{1}{1000} + \frac{1}{1000})$.

D'autres méthodes ont été proposées par la suite. Pritchard et al, [53] proposent une approche bayésienne de regroupement d'individus, implémentée dans STRUCTURE. En 2006-2007, des approches de la famille de statistiques multivariées, de type réduction de la dimension, ont été proposées, comme l'Analyse en Composante Principale (ACP) et le positionnement multidimensionnel (MDS). La première approche est implémentée dans EIGENSTRAT [54,55] et la deuxième est implémentée dans PLINK [14]. Nous allons montrer, ici, le principe général de ces approches.

STRUCTURE : Supposons un modèle dans lequel il y a *Z* populations inconnues. Chaque population se caractérise par des fréquences alléliques différentes des autres populations. L'objectif de cette approche est d'assigner les individus -en se basant sur leurs génotypes- aux différentes populations.

Soit *X* la matrice de génotypes des sujets, *Z* les différentes populations inconnues auxquelles appartiennent les sujets et enfin *P* les fréquences alléliques inconnues et propres à chaque population. Cette méthode suppose principalement l'équilibre de Hardy Weinberg et l'indépendance de liaison génétique entre les marqueurs. Sous ces hypothèses, les allèles aux locus sont indépendamment distribués selon la distribution allélique appropriée à chaque population. Par suite, les probabilités $Pr(X|Z,P)$ sont bien connues.

L'approche bayésienne adoptée ici définit des lois *a priori* pour Z et P : $Pr(Z)$ *et* $Pr(P)$. Elle incorpore l'incertitude dans l'estimation des paramètres et l'évaluation de l'évidence des

groupes ainsi formés. La distribution *a posteriori* de *Z* et *P* sachant les génotypes (i.e. *X*) est donnée par la loi suivante :

$$\Pr(Z, P|X) \propto P(X|Z, P) \times \Pr(Z) \times \Pr(P).$$

Analyse en Composante Principale (ACP): L'analyse en composante principale a été utilisée dans les études génétiques, il y a trente ans par Cavalli-Sforza [56]. Le principe de base de cette méthode est simple.

Soit *n*,*m* les dimensions de la matrice *X* des génotypes. Les lignes représentent les individus (*n*) et les colonnes représentent les génotypes des SNPs (*m*). La matrice *X* s'écrit comme suit :

$$X = \begin{bmatrix} x_{11} \dots x_{1j} \dots x_{1m} \\ \dots \dots \dots \dots \dots \dots \dots \\ x_{i1} \dots x_{ij} \dots x_{im} \\ \dots \dots \dots \dots \dots \dots \dots \\ x_{n1} \dots x_{nj} \dots x_{nm} \end{bmatrix} \text{ avec } x_{ij} = \text{le nombre d}'\text{allèle de référence (i. e. } 0, 1, 2).$$

Cette matrice doit être centrée et réduite en faisant le calcul suivant :

1) $\mu_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}$: est la moyenne de la colonne j;
2) $p_j = \frac{\mu_j}{2}$: est la MAF du SNP j ;
3) $x'_{ij} = \frac{x_{ij} - \mu_j}{\sqrt{p_j(1-p_j)}}$: représente les génotypes centrés réduits;
4) La matrice centrée réduite est donc:

$$X_{cr} = \begin{bmatrix} x'_{11} \dots x'_{1j} \dots x'_{1m} \\ \dots \dots \dots \dots \dots \dots \dots \\ x'_{i1} \dots x'_{ij} \dots x'_{im} \\ \dots \dots \dots \dots \dots \dots \dots \\ x'_{n1} \dots x'_{nj} \dots x'_{nm} \end{bmatrix}.$$

Après, on calcule la matrice de variance-covariance entre les individus:

$$V = \frac{1}{m} X_{cr} X'_{cr} \text{ avec } X'_{cr} \text{ est la matrice transposée de } X_{cr}.$$

La matrice *V* est une matrice carrée de dimensions *n*×*n*.

On calcule ensuite les valeurs propres et les vecteurs propres de la matrice *V*. Les vecteurs propres sont appelés les axes des composantes principales (PC). Ces axes vont séparer les différentes populations et vont déterminer les outliers. Les outliers sont des individus dont la

structure génétique s'écarte, en plusieurs unités d'écart type, de la structure génétique moyenne de l'échantillon. Cette approche est implémentée dans EIGENSTRAT [55].

Positionnement multidimensionnel (MDS) : C'est aussi une méthode statistique de regroupement. Elle consiste à calculer les valeurs propres et les vecteurs propres de la matrice de similarité entre individus et non pas la matrice de variance-covariance comme dans l'ACP. Les valeurs de la matrice de similarité sont les distances entre les individus, calculées en utilisant les identités par état (Identity By State, « IBS »). La distance entre deux individus *i* et *j* pour l'ensemble des génotypes des *m* SNPs est calculée comme suit:

$$DST_{ij} = 1 - \frac{IBS_2 + 0.5 \times IBS_1}{m},$$

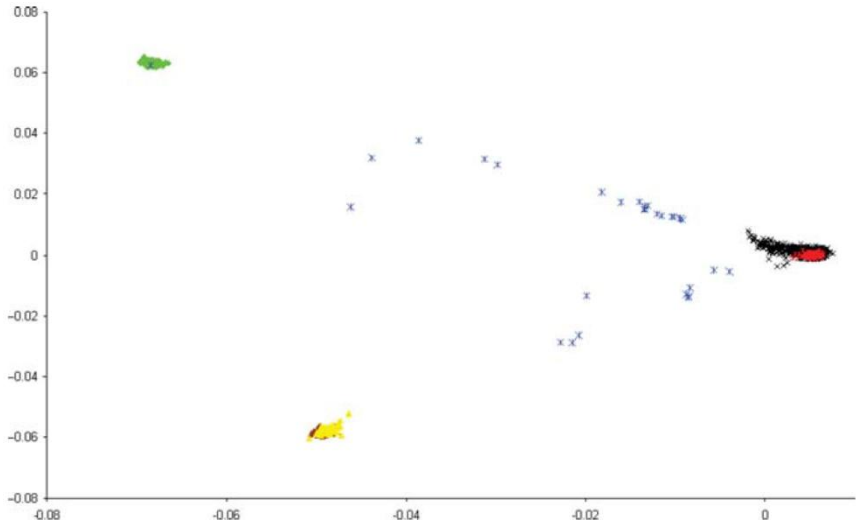
avec IBS_j, le nombre de SNPs où j allèles sont identiques chez les deux individus. Cette approche est implémentée dans PLINK [14].

Ces trois approches sont impactées par la présence de LD entre les SNPs. En effet, les axes estimés par l'ACP peuvent être perturbés par le LD. Le regroupement des individus pourra refléter le LD (surtout les grands blocs de LD) et non pas la stratification de population. Il faut donc que les SNPs utilisés soient indépendants. Pour mieux contraster les individus de l'étude, on peut ajouter à l'ACP des échantillons de sujets issus de diverses origines géographiques et connues. Typiquement, les données génotypiques des quatre populations de HapMap phase II et III (Yoruba, Ibadan-Nigeria ; Japonais, Tokyo-Japon ; Chinois, Pékin-Chine ; CEPH : résidents aux Etats Unis, originaires de l'Europe du Nord et de l'Europe de l'Ouest) peuvent être utilisées. Grâce à la diversité génétique de ces populations, deux axes principaux suffisent pour les séparer et identifier les outliers. Il faut noter que dans les études GWAS, qui contiennent des milliers de sujets et des centaines de milliers de SNPs, le calcul est intense et prend beaucoup de temps. Dans ce cas, la méthode implémentée dans STRUCTURE est quasi-impossible à réaliser. En revanche, l'ACP et le MDS sont plus rapides et faisables. En général, ces deux méthodes donnent les mêmes résultats et elles sont les plus utilisées dans les études d'association pan-génomiques.

Pour réduire le LD entre les SNPs dans notre étude GWAS de la MP, nous avons sélectionné un sous ensemble de 55193 SNPs « indépendants ». La procédure a été : un SNP est exclu si la mesure r^2 de LD est plus grande que 0.2 avec un autre SNPs dans la région de 1000 SNPs

consécutifs. En plus, nous avons exclu les régions contenant de longs intervalles de LD (chr2, chr5, chr6, chr8 et chr11) citées dans Price et al. [9]. Nous avons combiné nos données de SNPs aux mêmes SNPs extraits des données de 381 sujets indépendants issus des trois populations de HapMap (CEU, YRI et JPT-CHB). Nous avons appliqué l'ACP en utilisant EIGENSTRAT [54]. Les deux axes principaux séparent clairement les populations et regroupent les individus de notre GWAS (en noir) avec les CEU (en rouge) de HapMap (Figure 3.2). Ils identifient aussi 32 outliers (en bleu) que nous avons exclu de nos analyses. Finalement, le nombre total de sujets postQC était: 1039 cas et 1984 témoins.

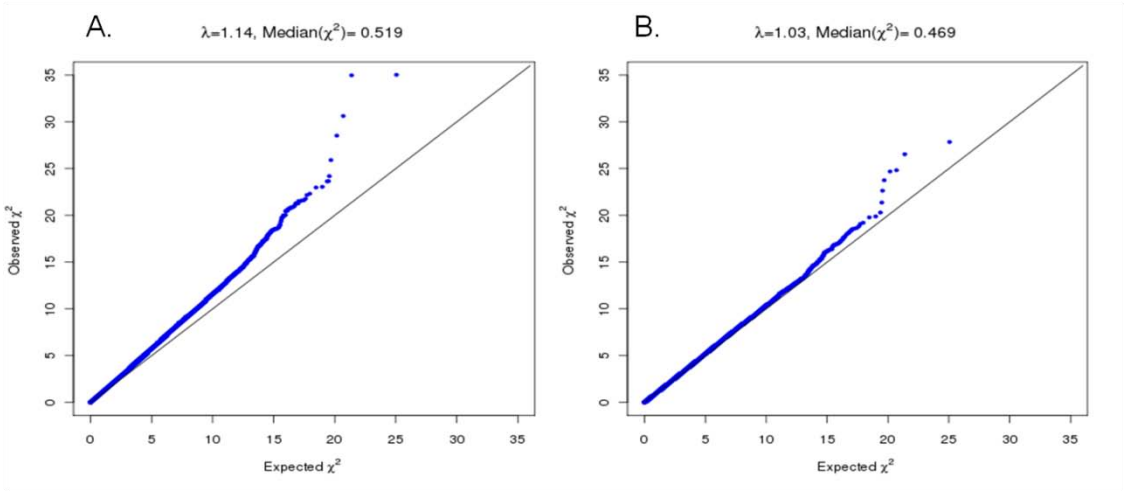
Figure 3.2- Analyses en composantes principales dans l'étape 1. La première composante est en abscisse et la deuxième composante est en ordonnée. Les individus sont représentés par des couleurs: rouge=CEU, jaune&marron =CHB&JPT, vert=YRI, noir=nos individus et bleu=outliers.



e/ Tests d'association

Dans l'étape 1, nous avons utilisé la régression logistique pour réaliser les 492929 tests d'association. Dans un premier temps, nous avons utilisé le modèle logistique sans covariables: $\text{Logit}(p) = \beta \times SNP$. Nous avons observé une légère inflation de la distribution du test d'association dans nos données (Figure 3.3 (A)) ($\lambda = 1.14$ et $\text{mediane}\ (\chi^2) = 0.519$).

Figure 3.3- Quantile-Quantile plot de la distribution de χ^2 de tests d'association : A. Modèle sans covariables et B. modèle avec deux covariables : PC1 et PC2.



Dans un deuxième temps, nous avons utilisé les deux axes principaux de l'ACP comme covariables dans le modèle logistique. Ici, le coefficient du contrôle génomique λ est égal à 1.03 (Figure 3.3 (B)) (médiane = 0.472).

Dans l'étape de réplication *in silico,* le modèle logistique sans covariables a été conduit parce qu'il n'y a pas de problème d'inflation de distribution dans les données du WTCCC2. Dans l'étape de réplication *de novo,* les individus provenaient de différentes origines géographiques : nous avons donc utilisé le test de Mantel-Haenszel Test pour tester l'association en tenant compte des centres géographiques, et le test de Breslow-Day pour tester l'hétérogénéité des ORs (expliqués dans le chapitre 1, section 1.2.2.1).

- **Résultats**

Deux SNPs ont été identifiés au niveau de signification pan-génomique (i.e. 0.05/492929 = 10^{-7}). Les deux SNPs appartiennent au gène SNCA, dont une mutation à transmission dominante explique certaines formes mendéliennes de la MP à âge de début précoce (chapitre 2, section 2.2). La signification des 50 meilleurs signaux d'association variait entre 5×10^{-5} et 2×10^{-8}. Ces 50 SNPs mappaient sur 23 locus indépendants : quatre SNPs dans SNCA, 11 SNPs dans MAPT, un SNP dans BST1 et 34 dans des locus non-connus pour la MP. Ces SNPs ont tous été testés pour la réplication *in silico* dans la cohorte anglaise. Dans cette étape, on déclare un SNP comme répliqué si sa valeur P est < 5% et son effet (i.e. β) a le même signe que dans l'étape du scan. Les 15 SNPs de SNCA ($P < 5\times10^{-5}$) et de MAPT ($P<10^{-6}$ pour MAPT) ont tous été fortement répliqués. Leurs fréquences alléliques chez les témoins sont similaires dans les deux étapes (Table 3.1). Le SNP du gène BST1 a aussi été répliqué mais

avec une évidence statistique moins forte (P= 0.02, OR =1.08). Parmi les 20 loci restants, quatre SNPs de trois loci ont été répliqués.

Table 3.1- Meilleurs résultats de l'étape 1 aux gènes SNCA, MAPT et BST1 et résultats correspondants dans la cohorte de réplication *in silico*

Chromosome (gene)	Position (bp)	SNP	Stage-1: scan (France) data					Stage-2: replication (UK) data		
			RA[a]	RAF[b]	OR	P_{GC} (two-tailed)[c]	P_{2PCs} (two-tailed)[d]	RAF	OR	P (one-tailed)[e]
Known PD genes/previously published loci										
4q22 (SNCA)	90858538	rs11931074	T	0.07	1.52	1.35E-05	9.04E-05	0.07	1.33	4.01E-05
	90860363	rs356220	T	0.35	1.37	2.82E-08	6.26E-07	0.36	1.27	2.59E-09
	90894261	rs3857059	G	0.07	1.54	1.00E-05	6.32E-05	0.07	1.33	3.95E-05
	90897564	rs2736990	G	0.44	1.35	2.88E-08	1.32E-07	0.45	1.24	3.98E-08
17q12–21 (MAPT)	41074926	rs393152	A	0.75	1.32	2.68E-05	1.43E-04	0.76	1.31	2.20E-08
	41279463	rs12185268	A	0.76	1.32	3.44E-05	1.62E-04	0.76	1.30	3.59E-08
	41279910	rs12373139	G	0.76	1.33	1.81E-05	7.60E-05	0.76	1.30	2.77E-08
	41281077	rs17690703	C	0.72	1.34	3.94E-06	6.61E-06	0.72	1.24	1.37E-06
	41347100	rs17563986	A	0.75	1.34	1.30E-05	5.65E-05	0.76	1.31	2.58E-08
	41412603	rs1981997	G	0.76	1.33	2.20E-05	8.81E-05	0.76	1.30	4.61E-08
	41436901	rs8070723	A	0.75	1.33	2.19E-05	8.91E-05	0.76	1.30	2.61E-08
	41544850	rs7225002	A	0.59	1.27	2.72E-05	4.08E-05	0.57	1.23	1.14E-07
	41602941	rs2532274	A	0.75	1.33	2.21E-05	1.06E-04	0.75	1.28	2.92E-07
	41605885	rs2532269	T	0.75	1.33	1.90E-05	8.58E-05	0.76	1.29	1.11E-07
	41648797	rs2668692	G	0.76	1.33	1.97E-05	8.20E-05	0.76	1.29	1.22E-07
4p15 (BST1)	15346446	rs4698412	A	0.52	1.28	6.88E-06	1.96E-06	0.55	1.08	0.0247

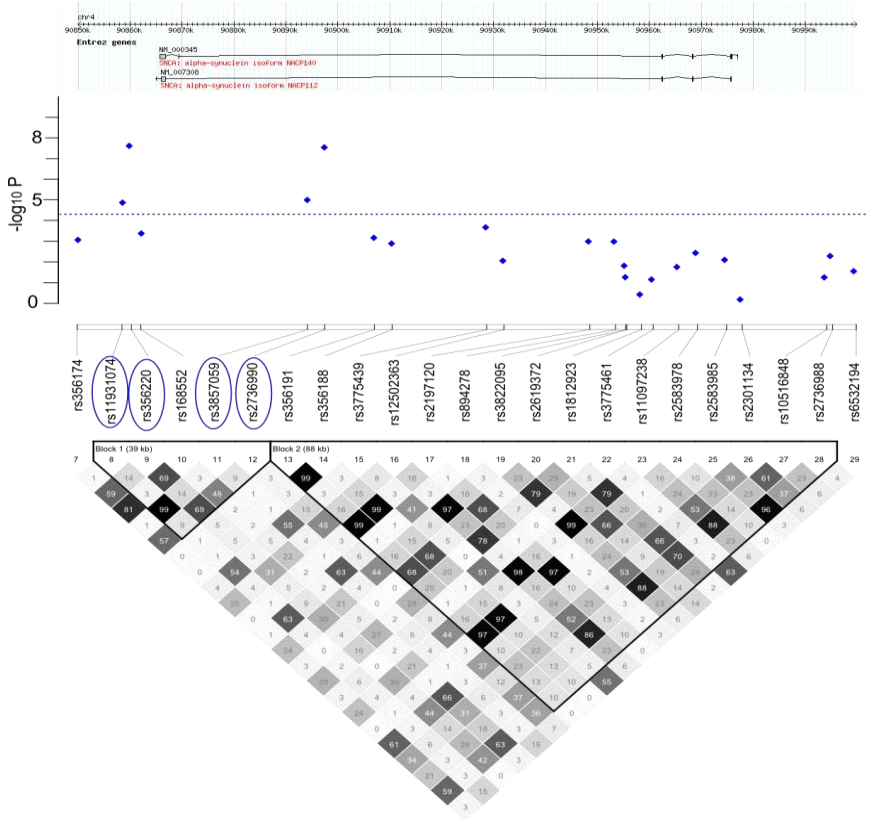
Dans l'étape de réplication *de novo*, nous avons testé ces quatre SNPs et le SNP de BST1. Deux SNPs ont été répliqués dans cette étape: le SNP de BST1 (P =0.029, OR=1.1) et un autre SNP (P=0.017, OR= 1.12) localisé sur le chromosome 12 (12q24) (Table 2 de l'article [57]).

Dans le gène SNCA, les SNPs rs11931074 et rs3857059 ont des niveaux de signification et des fréquences alléliques similaires. On peut donc se poser la question de savoir si ces deux signaux d'association sont indépendants ou non. La structure de LD dans le gène SNCA est montrée dans la figure 3.4. En fait, le LD entre ces deux SNPs est assez faible (r^2= 0.09).
Pour tester l'indépendance de ces deux SNPs, nous avons réalisé le modèle conditionnel suivant :

$$(1)\ \mathrm{Logit}(p) = \beta_1 \times \mathrm{rs2736990} + \beta_2 \times \mathrm{rs3857059} + \gamma \mathrm{PC}$$

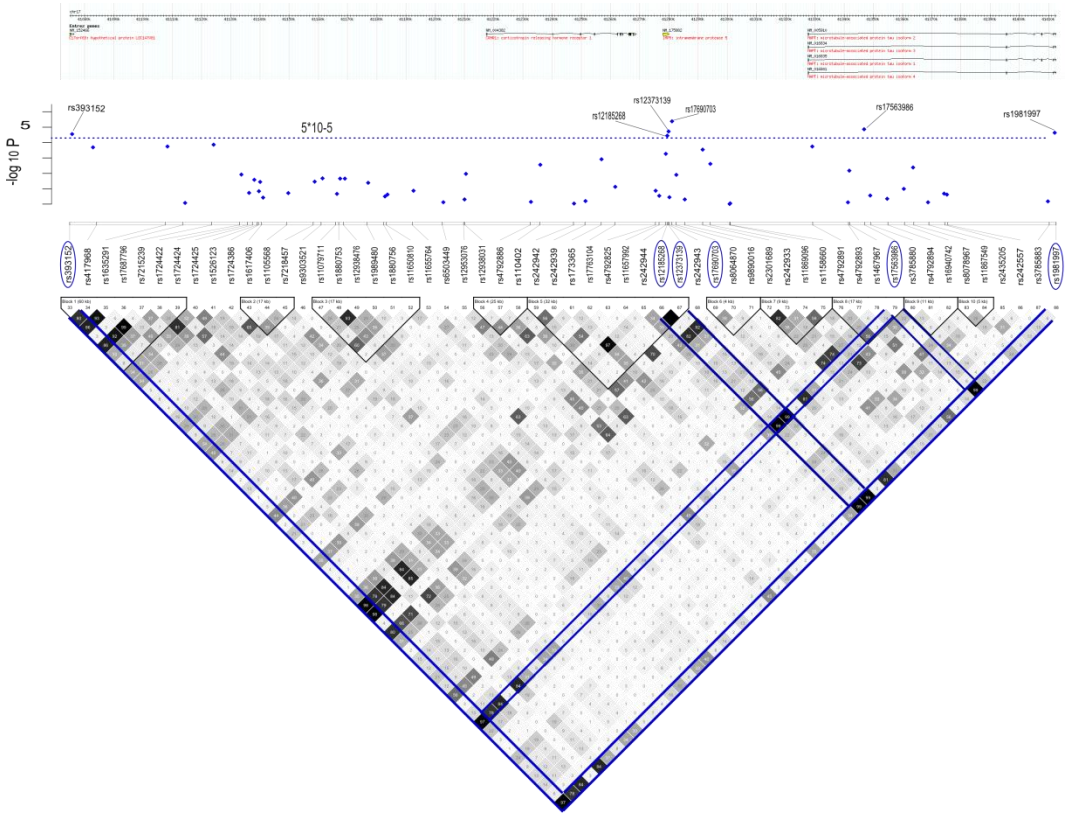
et nous avons testé l'hypothèse nulle $\beta_1 = 0$. La signification est 1.9×10^{-6}. Elle ne diminue pas beaucoup par rapport au résultat montré dans la table 3.1. Ces résultats suggéraient que le gène SNCA contient deux signaux relativement indépendants. En effet, ceci a été confirmé dans l'étude d'association pan-génomique de nos collaborateurs Anglais [41].

Figure 3.4- Structure de LD dans le gène SNCA et les résultats du test d'association



Dans le gène MAPT, les 11 SNPs associés représentent un seul signal d'association. Ce gène est situé dans une région de plusieurs longs blocs de LD (Figure 3.5). Deux études précédentes [58,59] ont identifié un large bloc de LD associé à la MP avec un haplotype à risque, appelé H1, et un haplotype protecteur, appelé H2. Les deux SNPs les plus significatifs (rs17690703 et rs17563986) dans notre étude sont localisés dans ce bloc de LD : le dernier SNP est localisé dans le gène MAPT. L'haplotype H2 est tagué par l'allèle mineur de quatre SNPs: rs12185268/G, rs12373139/A, rs1981997/A et rs8070723/G. Dans l'étape de scan, nous avons identifié le même signal d'association H1/H2 où les allèles mineurs des quatre SNPs sont significativement associés avec une diminution du risque de la MP (table 3.1, OR ~ 0.76 et $P<3.44 \times 10^{-5}$).

Figure 3.5- Structure de LD dans le gène MAPT et résultats du test d'association



Fonction des gènes BST1 et RFX4:

Le gène BST1 joue un rôle dans la génération des ADP-riboses cycliques qui sont des messagers pour la mobilisation du calcium (Ca^{2+}) dans le réticulum endoplasmique. L'interruption de l'homéostasie de Ca^{2+} a été récemment proposé comme une cause de la vulnérabilité des neurones dopaminergiques dans la MP [60].

La protéine RFX appartient à la famille "winged-helix" des facteurs de transcription "helix-turn-helix". Le transcrit RFX4_*v3* est le seul isoforme exprimé dans le cerveau uniquement. De plus, cet isoforme joue un rôle dans la transcription de plusieurs gènes impliqués dans la morphogenèse du cerveau.

Au total, les deux gènes BST1 et RFX4 sont fonctionnellement liés et indirectement impliqués dans la régulation de la concentration intracellulaire de Ca^{2+} qui joue un rôle assez important dans les fonctions et la mort des cellules.

En conclusion, notre étude d'association pan-génomique de la MP confirme l'implication de deux gènes, déjà connus comme impliqués dans la susceptibilité à la MP (SNCA et MAPT). Nous avons aussi confirmé, pour la première fois dans la population européenne, l'association avec BST1. Cette association a été confirmée par deux autres études ultérieures dans la population européenne [61,62]. Enfin, nous avons identifié un nouveau locus, 12q24. Ce locus se situe à 20kb du gène RFX4 qui contient des polymorphismes impliqués dans la maladie bipolaire [63]. Son rôle dans la MP reste non connu. En revanche, nous n'avons pas pu confirmer d'autres gènes connus de la MP, comme LRRK2, GBA ou PARK2.

Globalement, nos résultats et ceux des cinq autres GWASs de la MP publiées à ce jour, restent plutôt décevants. Toutes ces études n'ont clairement identifié que deux gènes SNCA et MAPT [41]. L'étude Hollandaise [61] a aussi confirmé le gène BST1. La troisième rapporte l'association avec *HLA-DRB5* [64]. Malgré les grandes tailles d'échantillons utilisés dans ces études et malgré l'utilisation des puces de génotypage à couverture assez importante (>80%), ces études pourraient individuellement manquer de puissance. Ceci a motivé la création, en 2010, du consortium IPDGC « International Parkinson's Disease Genomics Consortium » incluant des données pan-génomiques de six études : Nord-Américaine, Anglaise, Hollandaise, Allemande et Islandaise et la notre, Française. L'objectif principal de l'IPDGC est de maximiser la puissance pour détecter de nouveaux variants génétiques impliqués dans la MP. Il s'agissait de maximiser la puissance en combinant nos différents GWASs, et donc augmenter la taille d'échantillon, et en utilisant les analyses d'imputation afin de maximiser la couverture de la variabilité génétique sous-jacente.

Dans la section suivante, nous allons présenter notre étude de méta-analyse réalisée dans le consortium IPDGC.

3.1.2. Méta-analyse de cinq GWASs de l'IPDGC

Dans cette section, nous allons décrire les méthodes de méta-analyse que nous avons utilisées et le principe d'autres méthodes. Nous montrons ensuite l'analyse d'imputation. Nous terminons par l'exposé des résultats de la méta-analyse et la conclusion.

- **Méthodes statistiques de méta-analyse :**

Dans plusieurs domaines scientifiques, comme la biologie d'évolution, on cherchait à combiner les informations de plusieurs études afin d'augmenter la puissance statistique [65]. Pour arriver à cela, on peut idéalement grouper les données brutes de toutes les études et faire le test statistique sur les données ainsi groupées. Cependant, les données brutes ne sont souvent pas disponibles. Des approches de méta-analyse pour combiner les estimations des effets statistiques des études ont été proposées. Ces approches représentent un outil puissant, pouvant atteindre des puissances similaires à l'approche de regroupement des données brutes. Le principe de la méta-analyse consiste à estimer l'effet global de toutes les études.

Dans la méta-analyse de GWASs, on cherche souvent à estimer les effets combinés des SNPs. Lorsque les tailles d'échantillons des études sont égales, une façon simple de calculer l'effet global est de calculer la moyenne des effets individuels des études. Dans le cas contraire, il faut calculer la moyenne des effets individuels pondérés par la précision de chaque étude (fonction de l'erreur standard ou de la taille d'échantillon) afin de donner plus de poids aux résultats issus d'études basées sur des grands échantillons et moins de poids à celles basées sur des petits échantillons.

La première méthode de méta-analyse a été proposée par Fisher en 1932 :

$$Z_T = \sum_{i=1}^{k} -2 \times \ln(P_i) \sim \chi^2_{(k)}$$

Avec P_i est la valeur-p du test unilatéral de l'étude *i (i=1,…,k)* .

En 1949, Stouffer et ses collègues [66] ont proposé une méthode, proche de celle de Fisher, connue par le nom de *Stouffer* ou l'*Inverse Normal*. Elle consiste à combiner les Z-scores du test unilatéral plutôt que les valeurs-p :

$$Z_T = \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}} \sim N(0,1)$$

Avec $Z_i = \Phi^{-1}(P_i)$ et Φ^{-1} est l'inverse de la fonction de distribution de la loi normale centrée réduite et k est le nombre d'étude.

L'inconvénient majeur de ces deux approches est qu'elles ne tiennent pas compte de la différence des tailles d'échantillons des études. En 1958, Lipták [67] a proposé une méthode similaire à la méthode de Stouffer mais avec des poids. Cette méthode est connue par le nom de Z-test pondéré :

$$Z_T = \frac{\sum_{i=1}^{k} w_i Z_i}{\sqrt{\sum_{i=1}^{k} w_i^2}} \sim N(0,1)$$

Pour un test bilatéral, le Z score de chaque étude est calculé comme suit : $Z_i = \Phi^{-1}(P_i) \times \text{signe}(\text{Effet}_i)$ [68]. Plusieurs poids peuvent être utilisés comme la racine carré de la taille d'échantillon $w_i = \sqrt{N_i}$ ou l'erreur standard de l'effet.

Une autre méthode, appelée l'inverse de la variance est largement utilisée et nécessite l'utilisation des estimations des effets β_i et leurs erreurs standards SE_i. Cette méthode est représentée par les équations suivantes :

$$Z_T = \frac{\beta_T}{SE_T} \sim N(0,1)\ avec$$

$$\beta_T = \frac{\sum_{i=1}^{k} \beta_i \times w_i}{\sum_{i=1}^{k} w_i};$$

$$SE_T = \frac{1}{\sqrt{\sum_{i=1}^{k} w_i}};$$

$$w_i = \frac{1}{SE_i^2}.$$

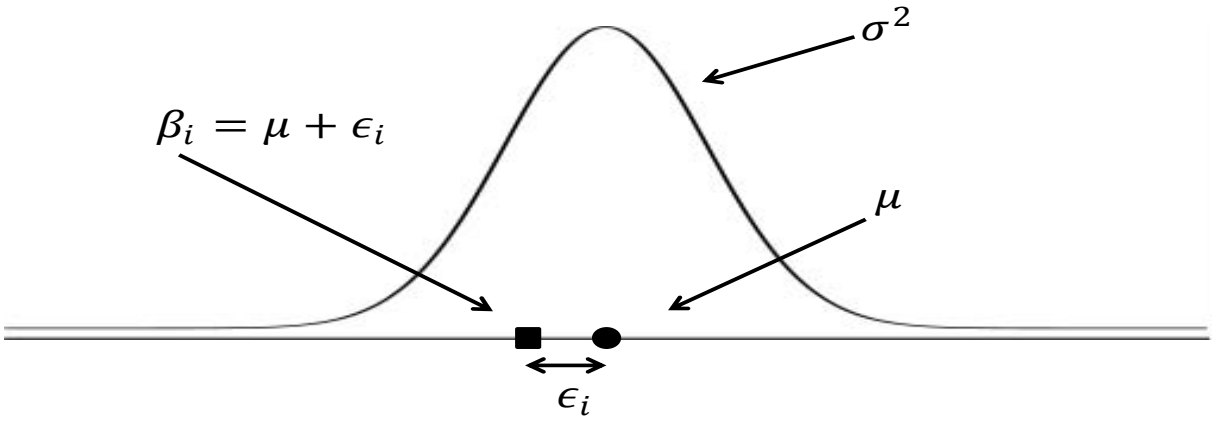
Asymptotiquement, les approches de méta-analyse de l'inverse de la variance et le Z-test pondéré par la racine carré de la taille d'échantillon sont similaires. En effet,

$$Z_{T(InvVar)} = \frac{\beta_T}{SE_T} = \frac{\sum_{i=1}^{k} \beta_i \times \frac{1}{SE_i^2}}{\sqrt{\sum_{i=1}^{k} \frac{1}{SE_i^2}}} = \frac{\sum_{i=1}^{k} \frac{\beta_i}{SE_i} \times \frac{1}{SE_i}}{\sqrt{\sum_{i=1}^{k} \frac{1}{SE_i^2}}} = \frac{\sum_{i=1}^{k} Z_i \times \frac{\sqrt{N_i}}{\sigma}}{\sqrt{\sum_{i=1}^{k} \frac{N_i}{\sigma^2}}}$$

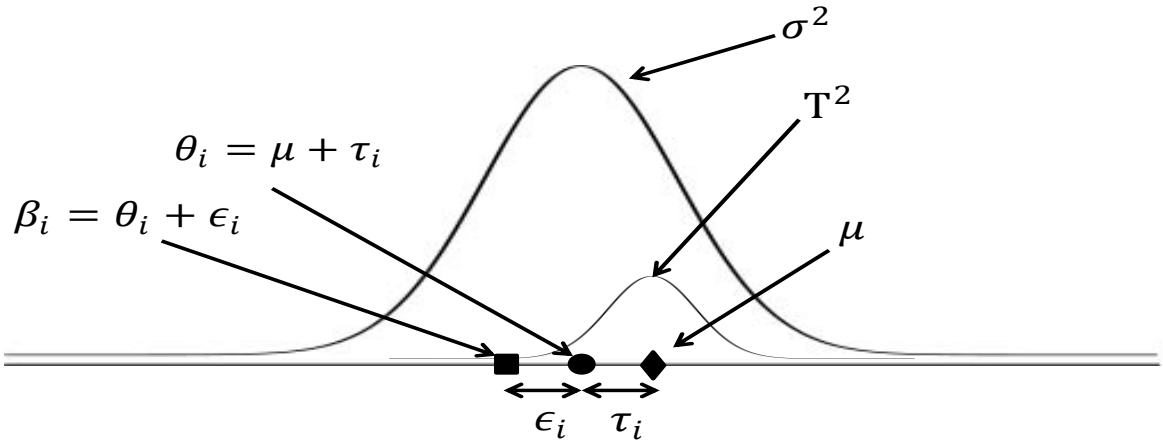
$$= \frac{\sum_{i=1}^{k} Z_i \times \sqrt{N_i}}{\sqrt{\sum_{i=1}^{k} N_i}} = \frac{\sum_{i=1}^{k} w_i Z_i}{\sqrt{\sum_{i=1}^{k} w_i^2}} = Z_{T(z-test\ pondéré)}$$

avec $w_i = \sqrt{N_i}\ et\ \sigma^2$ est la variance dans la population.

Dans les approches précédentes, le modèle supposé est un modèle à effet fixe. Il suppose que toutes les études partagent un effet commun μ. Les effets observés sont distribués selon une loi de moyenne μ et de variance σ^2 qui dépend essentiellement de la taille d'échantillon de chaque étude : $\beta_i = \mu + \epsilon_i$ avec ϵ_i est l'erreur intra-étude.



Lorsque les effets sont hétérogènes entre études, les modèles à effets fixes ne sont pas valides. Il faut donc utiliser des modèles à effets aléatoires qui supposent que les effets dans chaque étude sont distribués selon une loi de moyenne θ_i et de variance σ^2. En plus, θ_i est distribué selon une loi de moyenne μ et de variance T^2.



Les effets s'écrivent donc comme : $\beta_i = \mu + \epsilon_i + \tau_i$ avec ϵ_i est l'erreur intra-étude et τ_i est l'erreur inter-étude. Sous ce modèle, la méta-analyse se base sur la décomposition de la

variance totale en deux parties : la variance intra-étude σ^2 et la variance inter-étude T^2. La variance totale est représentée par la statistique *Q* de Cochrane :

$$Q = \sum_{i=1}^{k} w_i(\beta_i - \beta_T)^2,$$

avec β_T est l'effet total sous le modèle à effet fixe. Le nombre de degré de liberté *df* représente la variance attendue si les effets dans toutes les études sont les mêmes. Finalement, la valeur *Q* – *df*, transformée à l'échelle de la variance intra-étude, nous donne la variance inter-étude T^2 :

$$T^2 = \begin{cases} \dfrac{Q - df}{C} & \text{si } Q > df \\ 0 & \text{sinon} \end{cases}$$

avec $C = \sum w_i - \frac{\sum w_i^2}{\sum w_i}$.

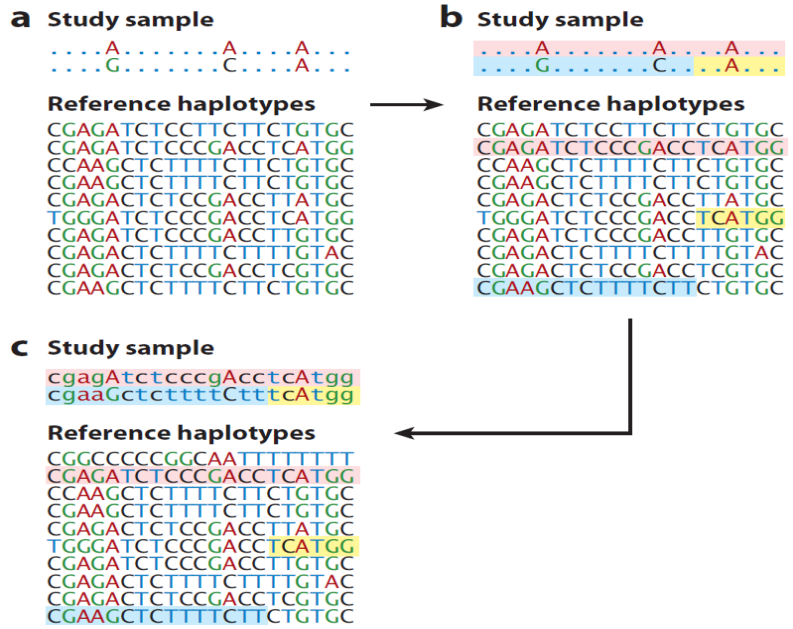
Pour calculer le Z-score globale Z_T, les équations du modèle de l'inverse de la variance restent les mêmes sauf que w_i est égale à $\frac{1}{SE_i^2 + T^2}$.

Dans notre étude de méta-analyse de la MP, nous avons utilisé la méthode de l'inverse de la variance à effet fixe. En cas d'hétérogénéité d'effets entre études (test de Cochrane), nous avons utilisé le modèle de méta-analyse à effet aléatoire.

- **Méthodes d'imputation:**

Qu'est ce que l'imputation? L'imputation est une technique d'inférence de génotypes manquants. Elle utilise 1) l'information génétique de ces génotypes combinés à des échantillons d'haplotypes d'une population de référence et 2) la variabilité génétique locale dans l'échantillon d'étude. Le principe d'imputation est représenté dans la figure 3.6.

Figure 3.6- Principe de l'imputation. Figure tirée des travaux de Li et al, [69].



L'imputation peut permettre d'améliorer la couverture génétique locale et ainsi augmenter la puissance de l'analyse d'association.

Plusieurs méthodes d'imputation ont été proposées et sont principalement basées sur des modèles de Chaines de Markov Cachées comme par exemple, MACH [70], IMPUTE [71]et BEAGLE [72]. Les probabilités des génotypes sont estimées par :

$$P_{ijk} = \Pr(G_{ij} = k), k \in \{0,1,2\}, \sum_{k} P_{ijk} = 1,$$

avec G_{ij} indique le génotype de l'individu *i* au SNP *j*.

L'estimation du nombre de l'allèle de référence est donné par: $D_{ij} = 2 \times P_{ij2} + 1 \times P_{ij1} + 0 \times P_{ij0}$. Les valeurs D_{ij} sont appelées « doses alléliques ».

Ces méthodes donnent des mesures pour évaluer l'incertitude d'estimation des génotypes manquants. Ces mesures sont basées généralement sur la variance des doses alléliques estimées, divisée par la variance des génotypes attendue sous l'équilibre de Hardy Weinberg. Ces mesures varient entre 0 et 1:

$$I = Var(D)/2p(1-p)$$

avec p est la fréquence du SNP dans la population d'haplotypes de référence. On peut interpréter l'indicateur I comme suit: l'information $I = \alpha$ dans un échantillon de N individus

indique que la quantité des données imputées au SNP est presque équivalente aux données génotypées de αN individus.

Cette information dépend de deux facteurs principaux: la taille d'échantillon d'étude et la taille d'échantillon des haplotypes de référence. Plusieurs échantillons publics de référence sont disponibles comme les données génotypiques de HapMap2 ou les données de séquençage du projet « 1000Genomes ».

L'indicateur I est le « $RSQR$ » pour MACH et le « $INFO$ » pour IMPUTE. Les auteurs de MACH définissent le $RSQR$ <0.3 comme seuil d'exclusion de variants mal-imputés. Cela correspond à $INFO$ <0.5 pour IMPUTE. Pour un SNP *i*, le *RSQR* de MACH peut s'écrire comme suit :

$$RSQR_i = \begin{cases} \dfrac{\sum_{j=1}^{N} D_{ij}^2 /N - \left(\sum_{j=1}^{N} D_{ij}/N\right)^2}{2 \times p_i(1-p_i)}, & si\ p_i \in]0,1[\\ 1, & si\ p_i = 0\ ou\ p_i = 1 \end{cases}$$

Est ce que ces critères varient par fréquence de SNPs ? Ces critères ont été proposés pour les variants fréquents. Pour les variants rares, des critères plus stricts doivent être utilisés à cause de la difficulté d'imputation de ces variants. En effet, l'information génétique locale dans l'échantillon étudié ne capture pas les variants rares, à cause du faible LD, à moins que la taille de l'échantillon de référence ne soit très importante.

Distribution du *RSQR* en fonction du MAF et de l'échantillon des haplotypes de référence

Pour montrer la relation entre le *RSQR* et le MAF d'une part et l'échantillon des haplotypes de référence d'autre part, nous avons imputé les génotypes manquants dans nos données pan-génomique de la MP en utilisant deux versions du projet 1000 Genomes : la version d'Août 2009 (112 haplotypes) et la version d'Août 2010 (566 haplotypes). Dans la table 3.2, nous montrons la distribution du *RSQR* par classe de MAF. Il est clair que la proportion de SNPs ayant une valeur de *RSQR*>0.3 augmente avec la classe de MAF. Ceci est vrai quelque soit la version du projet 1000 Genomes. Avec la nouvelle version, nous observons un plus grand nombre de variants rares imputés mais la proportion de SNPs passant le contrôle qualité (*RSQR*>0.3) n'est pas meilleure (Table 3.2).

Table 3.2: Distribution du nombre de SNPs imputés et le nombre de SNPs passant le QC (*RSQR*>0.3) par classe de MAF dans les données pan-génomiques françaises d'imputation selon la version du projet 1000 Genomes

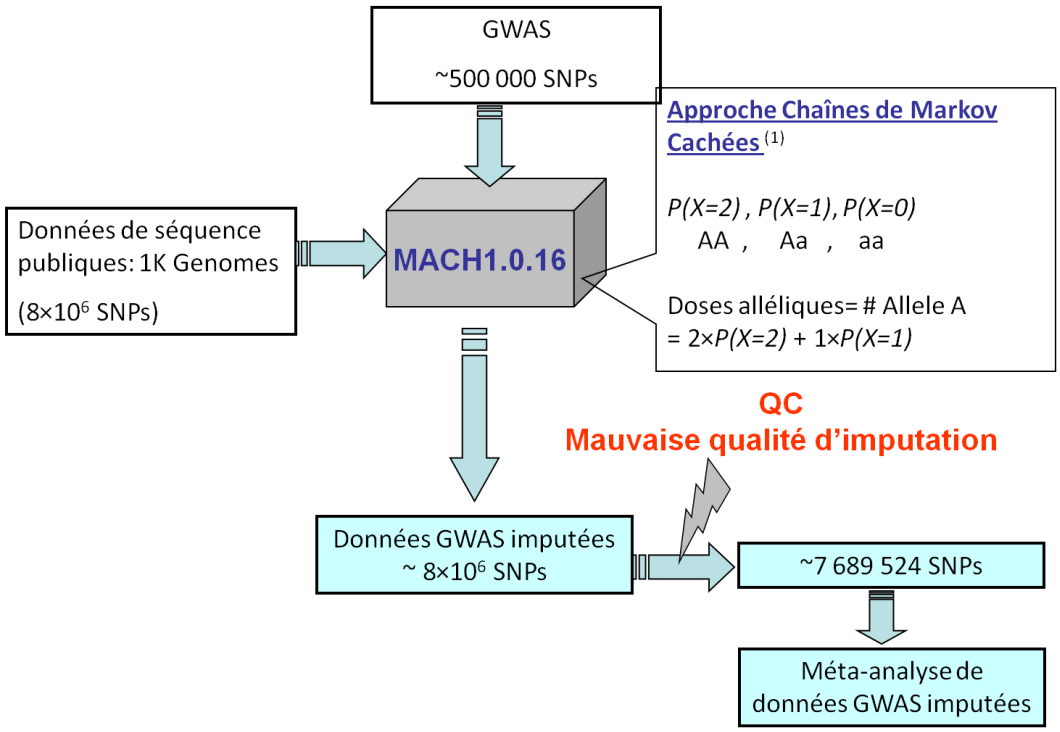
	1000 Genomes Août 2009				1000 Genomes Août 2010			
	Total imputé		RSQR>0.3		Total imputé		RSQR>0.3	
	#SNPs	(%)	#SNPs	(%)[$]	#SNPs	(%)	#SNPs	(%)[$]
[0 , 0.001[	482	0.01	187	38.8	1279677	11.06	138236	10.8
[0.001 , 0.005[	11983	0.15	8061	67.27	1785408	15.43	358684	20.09
[0.005 , 0.01[	63532	0.77	46989	73.96	814074	7.03	342124	42.03
[0.01 , 0.05[	1590698	19.35	1148355	72.19	2128316	18.39	1534388	72.09
[0.05 , 0.2[	3210065	39.05	2926770	91.17	2618441	22.63	2423317	92.55
[0.2 , 0.5]	3344314	40.67	3209677	95.97	2946585	25.46	2845125	96.56
Total	8221074	100	7340039		11572501	100	7641874	

[$] Pourcentage par rapport au #SNPs total imputé

Test d'association sur des données imputées : Le modèle de régression classique peut être utilisé pour tester l'association entre le trait Y et le SNP j: $Y_i = \alpha + \beta \times D_{ij}$. Ce modèle est implémenté dans plusieurs logiciels adaptés aux formats des sorties d'imputations. On distingue MACH2DAT/MACH2QTL (adaptés pour les sorties de MACH) et SNPTEST (adapté pour les sorties de IMPUTE).

Dans notre étude, nous avons utilisé la méthode MACH pour conduire les analyses d'imputations dans nos données de GWAS. Le design général de cette étude est schématisé dans la figure 3.7.

Figure 3.7- Design de la méta-analyse : Etape-1



L'imputation s'est déroulée en deux étapes: 1) estimer les paramètres de recombinaison dans un sous-échantillon de 200 individus combinés aux haplotypes de références et 2) inférer les génotypes dans le total des individus. Dans cette première méta-analyse nous avons utilisé la version d'août 2009 des données du 1000Genomes (112 haplotypes , 56 individus). Le test d'association a été réalisé pour les SNPs dont le RSQ était > 0.3, en utilisant le logiciel MACH2DAT. L'étude s'est déroulée en deux temps. Le premier consistait en une phase de scan dans un échantillon de l'IPDGC (5333 patients et 12019 témoins, (Table 3.3)). Après imputation, un total de 7689524 SNPs a été testé dans cette étape.

Table 3.3: Cohortes de deux étapes de la méta-analyse

	Etape 1- scan		**Etape 2- réplication**	
	Patients	**Témoins**	**Patients**	**Témoins**
USA	1850	3891	2807	2215
Angleterre	1705	5200	1271	1864
Allemagne	742	944	1153	712
France	1039	1984	267	363
Hollandais	-	-	1076	2426
Islandais	-	-	479	1427
Méta-analyse	**5333**	**12019**	**7053**	**9007**

- **Résultats**

Les analyses statistiques de cette méta-analyse ont révélé 11 variants génétiques associés à la MP avec des niveaux de signification pan-génomique ($P < 5\times10^{-8}$). Ces résultats ont été confirmés au cours de la phase de réplication dans un nouvel échantillon de 7053 patients et 9007 témoins (Table 3.3). Six des loci identifiés sont déjà connus (SNCA, MAPT, BST1, LRRK2, GAK et HLA-DRB5) et cinq autres sont nouveaux (ACMSD, STK39, MCCC1/LAMP3, SYT11, CCDC62/HIP1R).

Nous avons cherché à estimer l'impact des 11 variants identifiés sur la susceptibilité de la MP. Pour cela, nous avons d'une part calculé le risque combiné (Population Attributable Risk « PAR ») et d'autre part effectué des analyses de profils de risque.

Le PAR d'un SNP *j* dépend de sa fréquence et de son effet (i.e. OR) et se calcule comme suit :

$$PAR = \frac{MAF \times (OR - 1)}{MAF \times (OR - 1) + 1}$$

Le calcul de la valeur de PAR combiné de plusieurs SNPs se fait comme suit :

$$PAR_{combiné} = 1 - \prod_{j}(1 - PAR_j)$$

Le PAR combiné des 11 polymorphismes a été estimé à 60% dans l'étape de réplication. Cette mesure est souvent critiquée car les estimations des effets de SNPs sont probablement biaisées.

Les analyses de profils de risque consistent à calculer un score pour chaque individu *i* de l'étape de réplication, en se basant sur ses génotypes aux *N* SNPs identifiés dans l'étape du scan (*N*=11 dans notre cas): $\text{Score}_i = \sum_{j=1}^{N} \text{C}_{ij} \times \beta_j$; C_{ij} est le nombre d'allèles à risque de l'individu *i* au SNP *j* et β_j est l'estimation d'effet du SNP *j* dans l'étape 1. Le but de ces analyses est d'estimer la valeur prédictive des variants identifiés dans la première étape, à savoir, évaluer si les scores calculés chez les sujets de la deuxième étape permettent de les retrouver dans leur classe phénotypique (i.e. malade vs témoins). Pratiquement, on ordonne ces scores et on les classe dans cinq quintiles. Dans chaque quintile, on calcule la proportion des sujets malades relativement à celle obtenue dans le premier quintile. On procède de la même façon avec les sujets témoins, pour finalement obtenir les ORs à chaque quintile, en particulier celui du quintile le plus élevé (sujets ayant le plus grand nombres d'allèles à risques) par rapport au quintile le plus bas (sujets ayant le plus faible nombres d'allèles à risque). Dans notre étude, l'estimation de l'OR du 5ème quintile est 2.51 (IC=[2.23-2.83]) plus élevé que celui du 1er quintile (Table 3.4). Cette valeur est proche de l'estimation du risque de récurrence familial de la MP.

Table 3.4: Profiles de risques estimés dans les données de réplication de la méta-analyse [62]

Cohorte	**Quintile de risques (95%IC)**				
	1er (référence)	2ème	3ème	4ème	5ème
USA	1	1.49(1.25–1.78)	1.67(1.40–2.00)	1.9(1.59–2.27)	2.25(1.88–2.70)
UK	1	1.63(1.27–2.08)	2.26(1.77–2.88)	2.65(2.09–3.38)	3.3(2.60–4.21)
Allemand	1	1.16(0.86–1.57)	1.55(1.14–2.11)	1.68(1.23–2.29)	2.06(1.51–2.82)
France	1	1.24(0.72–2.16)	2.13(1.26–3.66)	2.84(1.68–4.88)	4.31(2.51–7.55)
Hollande	1	1.21(0.74–2.00)	1.12(0.68–1.84)	1.5(0.93–2.42)	1.89(1.17–3.07)
Méta-analyse	1	1.43(1.27–1.62)	1.77(1.55–1.99)	2.03(1.80–2.32)	2.51(2.23–2.83)
Cas (%)	886(0.39)	1069(0.4713)	1185(0.5216)	1268(0.5593)	1394(0.6117)

Analyse de l'expression des gènes

Pour comprendre les effets biologiques des variants identifiés, une analyse de l'expression des gènes sur des prélèvements post-mortem du cerveau de plus de 350 donneurs, neurologiquement saints au moment de la mort a été effectuée. Cette étude consiste à tester l'association entre les SNPs, situés dans les loci identifiés ± 1Mb en amont et en aval, et les expressions des gènes localisés dans ces loci. Le modèle utilisé est un modèle de régression linaire où la variable à expliquer est l'expression du gène et la variable explicative est le génotype du SNP. Des associations significatives (valeur-P < 3.55×10^{-5}) ont été trouvées pour cinq des 11 loci identifiés. Ce type d'analyse mérite plus d'investigation dans le futur pour élucider les fonctions biologiques sous-jacentes des variants d'ADN identifiés.

Conclusions générales

Notre étude de méta-analyse a été poursuivie par une nouvelle méta-analyse des données de l'IPDGC [73] adossée à un autre échantillon indépendant provenant du 23andMe (3426 cas et 29624 témoins) [74]. Au total, nos deux méta-analyses de la MP ont identifié près d'une vingtaine de locus associés au risque de la MP (Table 3.5). La grande majorité de ces locus sont nouveaux. Il y a trois ans, leur implication n'était pas connue dans les formes idiopathiques de la MP. Les études pan-génomiques confirment aussi l'existence d'un continuum entre les formes monogéniques et communes de la MP. Cependant, les études ne permettent pas d'identifier les variants expliquant la variation du risque de la MP, que ce soit ceux des gènes déjà connus de la MP ou les nouveaux.

Table 3.5: Loci de MP identifiés par étude d'association pan-génomique de nos deux méta-analyses [62,73]

Chr	Locus	Gène(s)	Pathologies
4	PARK4	GAK	MP dominante
4	PARK1	SNCA	MP monogénique dominante
12	PARK8	LRRK2	MP dominante
17		MAPT/STH	Démence Fronto-temporale (Tauopathie)
1	1q21	SYT11/RAB25/RIT1	
1	1q32	RAB7L1/PARK16	
2	2q21	ACMSD	
2	2q24	STK39	Autisme
3	3q26	NMD3	
3	3q27	MCCC1/LAMP3	
4	4p15	BST1	
4	4q21	STBD1	
6	6q21	HLA-DRB5	
7	7p15	GPNMB	
8	8p22	FGF20	
8	8q21	MMP16	
12	12q24	CCDC62/HIP1R	
16	16p11	STX1B	

3.2 Tests multi-marqueur : Etudes « gene-wide »

Nous venons de voir que les loci identifiés par nos analyses pan-génomiques n'expliquent qu'une faible part du risque génétique de la MP. Plusieurs hypothèses sont postulées dont celle de l'association indirecte, c.à.d. le SNP testé n'est pas le variant causal. L'estimation de l'effet est donc biaisée et, peut être sous-estimée. Par ailleurs, d'autres variants de la MP peuvent exister. Le taux de faux négatifs de nos analyses précédentes, même celles conduites au sein de l'IPDGC (> 20000 sujets) n'est pas à exclure. Plus généralement, on peut se poser la question de la variabilité génétique qui a été explorée par ces analyses. En effet, (i) un ou plusieurs variants communs de la MP peuvent se situer dans une région mal couverte par les puces de génotypage, même si ces puces couvrent en moyenne 90% de la variabilité génétique commune; (ii) certains variants de la MP ne sont pas communs, peu fréquents: ils ne sont donc ni observés ni fortement corrélés aux SNPs de la puce.

Pour augmenter la probabilité de l'association directe, il faut accéder à l'ensemble de la variabilité génétique et/ou maximiser l'information apportée par les marqueurs. Cette dernière stratégie repose sur les tests d'association multi-marqueur qui permettent d'analyser, contrairement au test simple-marqueur, plusieurs marqueurs conjointement.

Les tests multi-marqueur classiques sont ceux basés sur l'analyse des haplotypes : les tests haplotypiques, et ceux basés sur l'analyse jointe des génotypes de plusieurs marqueurs, les tests multivariés. Ces tests analysent les SNPs au sein d'une même unité génomique. Dans les études pan-génomiques, l'unité d'analyse peut être une fenêtre du génome de longueur fixée : par exemple, le génome est découpé en fenêtres de p SNPs ou x kilobases ; ces fenêtres peuvent être chevauchantes ou non. Plusieurs découpages du génome sont possibles selon le choix sur la longueur (p ou x). Une autre unité d'échantillonnage plus naturelle est celle du gène. L'inconvénient majeur de ces tests est l'accroissement du nombre de degrés de liberté du test avec le nombre de SNPs de l'unité d'analyse. Il a été rapporté que le gain d'information, utilisée par ces tests multi-marqueur, ne compense généralement pas l'augmentation du *df*, par rapport au test simple-marqueur [75].

D'autres tests d'association multi-marqueur ne dépendent pas d'un grand nombre de *df*. On distingue:

1- Les analyses en composantes principales qui résument l'information des SNPs par les composantes principales et réduisent donc le nombre de *df*;

2- Les méthodes de noyaux, à un seul *df*, qui résument l'information aux SNPs en une seule variable, la similarité génétique entre les individus.

Dans notre travail de thèse, nous avons exploré l'intérêt des méthodes de noyaux pour l'analyse d'association multi-marqueur.

3.2.1 Tests basés sur le modèle de noyau: Test « SNP–Set »

Schaid et al, [75] ont proposé un test non-paramétrique de noyaux pour les traits binaires. Les limites de ce test sont qu'il suppose le même signe des effets des SNPs, et ne permet pas de prendre en compte les effets de covariables, ni donc d'ajuster pour la stratification de population.

Plus récemment, Kwee et al, [76] ont proposé une méthode de régression non-paramétrique pour des traits quantitatifs. Cette méthode a été généralisée pour les traits binaires par Wu et al, [77]. Cette méthode se libère des deux inconvénients précédents. En effet, en proposant un test basé sur la cumulation des carrés des effets, cette méthode ne suppose pas un même signe des effets. De plus, par sa correspondance avec les modèles linéaires mixtes, elle permet de conduire le test d'association en ajustant pour des effets de covariables. Ces tests sont dénotés par les auteurs « SNP-Set association test ».

Nous allons d'abord décrire les aspects théoriques de ces tests. Nous les présentons historiquement. Ainsi, nous montrons des correspondances mathématiques entre certains de ces tests et méthodes.

Modèle général et notations

Notons par $X_i = (X_{i1}, \dots, X_{ij}, \dots, X_{ik})'$ le vecteur de génotypes de l'individus *i* aux *k* SNPs (*j=1,…,k*). Les génotypes X_{ij} sont codés en 0, 1 ou 2 (nombre de copies de l'allèle mineur). Notons par $Y = (Y_1, \dots, Y_i, \dots, Y_n)'$ les phénotypes des *n* sujets ($Y_i = 0$ pour les témoins et 1 pour les cas). Ainsi, notons par n_0 le nombre de témoins, par n_1 le nombre de cas et par *n* le nombre total d'individus.

Pour les traits binaires, le modèle logistique multivarié s'écrit sous la forme :

$$\text{Logit Pr}(Y_i) = \beta_0 + \sum_{j=1}^{k} X_{ij}\beta_j \qquad (1)$$

L'hypothèse nulle d'absence d'association est $H_0: \beta = (\beta_1, \dots, \beta_j, \dots, \beta_k)' = 0$ avec β_j est le coefficient de régression du SNP j. Les coefficients de régression sont estimés par le maximum de vraisemblance. Certains tests statistiques sont proposés pour tester l'hypothèse nulle : le test de Wald, le test de rapport de vraisemblance (LRT) ou le test de score. Dans ce qui suit, nous allons montrer le test de score, simple à calculer.

Dans le modèle (1) ci-dessus, le vecteur des scores et sa matrice de variance-covariance s'écrivent comme suit :

$$U = \sum_{i=1}^{n} (Y_i - \bar{Y}) X_i, \qquad \text{(vecteur de dimensions } k \times 1)$$

$$V = \bar{Y}(1 - \bar{Y}) \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})' \qquad \text{(matrice de dimensions } k \times k)$$

Avec $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$ et $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ les moyennes des phénotypes (proportions des cas) et des génotypes (2×MAF), respectivement.

Le test de score multivarié s'écrit sous la forme: $T^2_{score} = U'V^{-1}U$

qui suit, asymptotiquement, une distribution de χ^2 à $k - 1$ *df*.

Deux inconvénients majeurs : nombre de *df* et LD entre SNPs

D'une part, lorsque le *df* est assez grand, le test n'a pas une bonne puissance. D'autre part, le modèle peut être sur-paramétrisé si le niveau de LD entre les SNPs est fort.

TEST1 : Wang et Elston : réduction du LD et regroupement des SNPs en une méga-variable

Pour éviter ces deux problèmes, Wang et Elston [78] ont proposé un test de sommes pondérés. Une première étape consiste à réduire le LD entre les SNPs en transformant les génotypes en des composantes de Fourrier. La deuxième étape consiste à réduire le nombre de *df* en proposant le test de somme de scores pondérés (WST) suivant :

$$T_{score} = \frac{w'U}{\sqrt{w'Vw}} \sim N(0,1),$$

avec $w = (w_1, \dots, w_j, \dots, w_k)'$ le vecteur des poids ($w_j = \left[\frac{1}{j+1}\right]^2$). Le calcul du vecteur *U* et de la matrice *V* est le même que dans les équations du modèle (1). La seule différence est la variable X_i qui représente les composantes de fourrier et non pas les génotypes.

Dans ce test de score, le modèle (1) peut être vu comme un modèle à un *df* qui s'écrit comme suit :

$$\text{Logit Pr}(Y_i) = \beta_0 + \beta_c \sum_{j=1}^{k} X_{ij} \qquad (2)$$

L'hypothèse nulle est $H_0: \beta_c = 0$. Le coefficient de régression β_c peut être déduit des coefficients β_j des SNPs testés individuellement. La relation entre β_c et β_j a été montré pour la régression linéaire dans Pan [79] :

$$\beta_c = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n} X_{ij}^2 \beta_j}{\sum_{i=1}^{n}\left(\sum_{j=1}^{k} X_{ij}\right)^2}.$$

Le coefficient β_c est vu comme la moyenne pondérée des coefficients individuels.

Réduction du LD, *df*=1, mais l'effet global s'annule si les effets individuels n'ont pas les mêmes signes !

Lorsque les β_j ont des sens opposés, l'effet global a tendance à s'annuler. Il en résulte donc une perte de puissance. Pan [79] a proposé deux nouveaux tests. Le premier est basé sur les estimations des β_j des tests individuels et le deuxième est basé sur le test statistique de score *U*.

TEST2 : Test basé sur les coefficients de régression individuel $\boldsymbol{\beta_j}$

Version sans poids

Une première version de ce test est construite sans utiliser de poids. Le test s'écrit comme suit :

$$SumSqB = \beta'\beta = \sum_{j=1}^{k} \beta_j^2$$

Version avec poids

La deuxième version pondère les β_j par l'inverse de leurs variances V_{jj}^{-1} :

$$SumSqBw = \beta' diag(V^{-1})\beta = \sum_{j=1}^{k} \beta_j^2 / V_{jj}$$

(β_j et V_{jj} sont les coefficients de régression et leurs variances estimés dans le modèle simple-marqueur).

Signification des deux tests : formes quadratiques

Les deux tests $SumSqB$ et $SumSqBw$ ont la forme quadratique suivante : $Q = \beta' W^{-1} \beta$

avec $W = I$ dans le premier test et $W = diag(V)$ dans le deuxième test.

Il est connu que la distribution de Q suit une somme pondéré de *k* distributions de χ^2 à un df : $\sum_{j=1}^{k} c_j \chi_1^2$ [80], avec c_j les valeurs propres de la matrice VW^{-1}. Zhang JT [81] propose d'approximer cette distribution par la distribution suivante : $a\chi_d^2 + b$ avec :

$$a = \frac{\sum_{j=1}^{k} c_j^3}{\sum_{j=1}^{k} c_j^2}, b = \sum_{j=1}^{k} c_j - \frac{\left(\sum_{j=1}^{k} c_j^2\right)^2}{\sum_{j=1}^{k} c_j^3}, d = \frac{\left(\sum_{j=1}^{k} c_j^2\right)^3}{\left(\sum_{j=1}^{k} c_j^3\right)^2}.$$

Dans le même principe des tests $SumSqB$ et $SumSqBw$, les tests basés sur la statistique de score *U* sont proposés. Ces deux types de tests ont la même performance. L'avantage du deuxième est son lien avec d'autres méthodes comme le « Kernel Machine Regression », comme nous allons le voir.

TEST3 : Test basé sur le test statistique de score *U*

Version sans poids

Le test est basé sur la somme des carrés des scores *U* :

$$SumSqU = U'U = (Y - \bar{Y})'XX'(Y - \bar{Y})'.$$

Version avec poids

La forme pondérée de ce test est :

$$SumSqUw = U'Diag\left(I_f\right)^{-1} U \text{ avec } I_f = V = cov(U) = \bar{Y}(Y - \bar{Y})'(X - \bar{X})'(X - \bar{X}).$$

La signification des tests $SumSqU$ peut être calculée comme celle des tests $SumSqB$.

Comment le problème de la différence de signes des effets est-il évité ?

Reprenons le modèle (2). Pour tester l'hypothèse nulle H_0: $\beta_c = 0$, le test de score U_c s'écrit comme suit : $U_c = \sum_{j=1}^{k} U_j$

Ce test est sensible par rapport au signe de *U* et donc par rapport au signe de β_j. En revanche, le test $SumSqU$ s'écrit en fonction des carrés des U_j: $SumSqU = \sum_{j=1}^{k} U_j^2$. Ce test est donc invariant avec le signe de U_j et de β_j.

D'autres méthodes sont proposées comme le test de Goeman et le « Kernel Machine Regression » (KMR).
La première approche est basée sur le modèle linéaire mixte alors que la deuxième est basée sur la régression non-paramétrique. Bien que le modèle de ces méthodes diffère, les tests proposés sont assez liés. En effet, le modèle KMR peut être vu comme un modèle linéaire mixte. En plus, les tests KMR et celui de Goeman sont liés aux tests *SumSqU* décrits ci-dessus.

1. Connexion entre le test *SumSqU* et le test de Goeman

TEST4 : Qu'est ce que le test de Goeman ?

Goeman et al [82] ont proposé une méthode bayesienne pour tester un grand nombre de paramètres comme dans le cas de la régression multivariée. Le test d'association proposé revient à un test de score pour des effets aléatoires dans un modèle de régression multivariée. Les β_j du modèle (1) ne sont plus considérés comme des effets fixes mais plutôt comme des effets aléatoires qui suivent une loi arbitraire de moyenne $E(\beta) = 0$ et de $cov(\beta) = \tau\Sigma$.
Le test de l'hypothèse nulle H_0: $\beta = (\beta_1, \dots, \beta_j, \dots, \beta_k)' = 0$ revient au test H_0: $\tau = 0$. Sous cette hypothèse, le test de score de Goeman est : $S = \frac{1}{2} U' \Sigma U - \frac{1}{2} \mathrm{tr}(\Sigma V)$,

avec $\mathrm{tr}(\Sigma V)$ est la trace de la matrice ΣV. La signification de ce test est évaluée empiriquement par permutation des phénotypes.

Lien avec *SumSqU*

On peut voir que la matrice ΣV est invariable avec la permutation. L'utilisation de *S* revient donc à utiliser $S_p = U'\Sigma U$ qui est équivalent au test de score multivarié à *k df* si $\Sigma = V^{-1}$, au test *SumSqU* si $\Sigma = I^{-1}$ et au test *SumSqUw* si $\Sigma = \big(Diag(V)\big)^{-1}$.

2. Connexion entre le test *SumSqU* et le KMR : « SNP-Set » Test

TEST5 : Qu'est ce que le « Kernel Machine Regression » (KMR) ?

Les KMRs considèrent un modèle de régression non paramétrique qui s'écrit comme suit :

$$\text{Logit Pr}(Y_i) = \beta_0 + h(X_{i1}, \dots, X_{ik}) \qquad (3)$$

où $h(.)$ est une fonction de noyau non paramétrique, inconnue et à estimer [83]. Cette fonction offre une certaine flexibilité pour modéliser les effets des SNPs sur le trait. Ces fonctions de noyau mesurent la similarité génétique entre les individus.

Deux fonctions de noyau très populaires sont la fonction de noyau polynomial de degré d : $K(X_i, X_j) = (X_i' X_j + \rho)^d$ et la fonction de noyau gaussien : $K(X_i, X_j) = \exp\{-\frac{||X_i - X_j||^2}{\rho^2}\}$ avec $||X_i - X_j||^2 = \sum_{j=1}^{k}(X_{ik} - X_{jk})^2$ et ρ un paramètre inconnu [83].

D'autres fonctions de noyau sont communément utilisées dans la génétique quantitative [83,84]:

- Fonctions linéaires, $K(X_i, X_j) = \sum_{j=1}^{k} X_{ik} X_{jk}$.
- Fonctions d'identité par état (IBS), $K(X_i, X_j) = \sum_{j=1}^{k} IBS(X_{ik}, X_{jk})$ qui est égale à $\sum_{j=1}^{k}(2 - |X_{ik} - X_{jk}|)$ si le codage de génotypes est additif (0,1 ou 2).
- Fonctions quadratiques, $K(X_i, X_j) = (1 + \sum_{j=1}^{k} X_{ik} X_{jk})^2$

Le choix de la fonction de noyau est crucial et dépend de la question posée et du modèle génétique sous-jacent. Nous discutons de ce choix dans une section suivante.

Le but du modèle non paramétrique (3) est de résumer toutes les variables explicatives (SNPs) du modèle par une seule variable h. De cette façon, le nombre de *df* est réduit à 1. Par le théorème de représentant de Kimeldorf et Wahba, 1971, nous avons:

$$h_i = h(X_i) = \sum_{j=1}^{n} \gamma_j K(X_i, X_j)$$

où γ_j sont certains paramètres connus.

Le test de l'hypothèse nulle d'absence d'association entre le trait et les SNPs revient à tester l'hypothèse suivante: $h = (h(X_1), \ldots, h(X_n)) = 0$.

Notons *K* la matrice $n \times n$ tel que l'élément *(i,j)* correspond à $K(X_i, X_j)$, et $\gamma = (\gamma_1, \ldots, \gamma_n)$. Nous avons donc $h = \gamma K$. En traitant h comme une variable d'effets aléatoires, spécifiques à chaque sujet, de moyenne 0 et de matrice de variance-covariance τK, tester H_0: $h=0$ revient à tester H_0: $\tau = 0$. Le test statistique de score de la composante de la variance est:

$$Q = (Y - \bar{Y})' K (Y - \bar{Y})$$

qui est une forme quadratique qui suit asymptotiquement un mélange de loi de χ^2 et qui peut être approximée par une distribution de χ^2 de la forme suivante $\kappa\chi^2_\upsilon$ [81].

KMR et modèle linéaire mixte

Le modèle non-paramétrique (3) peut être vu comme un modèle logistique mixte. En effet, si la fonction de noyau est linéaire $K(X_i, X_j) = \sum_{j=1}^{k} X_{ik} X_{jk}$, le modèle (3) revient implicitement au modèle (1)

$$\text{Logit Pr}(Y_i) = \beta_0 + \sum_{j=1}^{k} X_{ij}\beta_j$$

avec β_j représente les effets aléatoires des SNPs qui suivent une loi arbitraire de moyenne zéro et de variance τ [84].

Lien avec *SumSqU*

Il est clair que si *K(.,.)* est une fonction de noyau linéaire, la matrice *K* défini précédemment est égale au produit matricielle *XX'* (*X* est la matrice des génotypes). Le test issu du KMR peut donc s'écrire:

$$\begin{aligned} Q &= (Y - \bar{Y})' XX' (Y - \bar{Y}) \\ &= SumSqSU \end{aligned}$$

Pour la version pondérée de *SumSqSUw*, il suffit de prendre une fonction de noyau linéaire pondérée $K(X_i, X_j) = \sum_{j=1}^{k} w_j X_{ik} X_{jk}$. La matrice *K* s'écrit donc comme le produit matriciel *XWX'* avec *W* est une matrice diagonale $(k \times k)$ des poids des *k* SNPs. On a donc:

$$Q = (Y - \bar{Y})'XWX'(Y - \bar{Y})$$
$$= SumSqSUw$$

Dernière remarque : Le test « SNP-Set » et les variants rares

Récemment, le test « SNP-Set » a été proposé dans le contexte d'analyse de variants rares. L'extension apportée dans cette dernière publication est l'introduction dans la fonction de noyau de poids sur les contributions individuelles des SNPs. Le nom donné par les auteurs à cette extension est SKAT [84]. Une autre méthode, C-alpha [85], a aussi été proposée récemment. Ici, nous décrivons la méthode C-alpha telle qu'elle est proposée par les auteurs. Puis, nous montrons son lien avec SKAT.

C-alpha : test proposé premièrement par Neyman et Scott [86] en 1966

Rappelons quelques notations: n_1 est le nombre de cas, n_0 est le nombre de témoins et $n = n_1 + n_0$ est le nombre total de sujets. Notons par $n_{j,(j=1,\ldots,k)}$ le nombre de sujets qui portent l'allèle rare au SNP j parmi les n sujets. Notons par m_k le nombre de cas qui portent l'allèle rare au SNP j. On suppose que $m_j \sim \text{Bin}(n_j, p_j)$ avec $p_j = p_0 = 1 - \frac{n_0}{n}$ sous l'hypothèse nulle de non-association (p_0 =proportion des cas = $\frac{1}{2}$ si le nombre de cas est égal au nombre de témoins et l'allèle rare peut être chez les cas et chez les témoins aléatoirement). Le test de C-alpha est basé sur les quantités suivantes:

$$T_C = \sum_{j=1}^{k} T_{C.j} = \sum_{j=1}^{k} \left(m_j - n_j p_0\right)^2 - n_j p_0 (1 - p_0)$$

$$V_C = \sum_{j=1}^{k} Var(T_{C.j}) = \sum_{j=1}^{k} E\left[\left(m_j - n_j p_0\right)^2 - n_j p_0 (1 - p_0)\right]^2$$

avec

$$Var\left(T_{C.j}\right) = \sum_{u=0}^{n_j} \left[\left(m_j - n_j p_0\right)^2 - n_j p_0 (1 - p_0)\right]^2 \times f(u|n_j, p_0)$$

et $f\left(u\middle|n_j, p_0\right) = \mathrm{C}\left(n_j, u\right) p_0^u (1 - p_0)^{n_j - u}$ est la probabilité Pr ($U = u$) pour $U \sim \text{Bin}\left(n_j, p_0\right)$.

Si tous les m_j sont indépendants, alors $Z = T_C / \sqrt{V_C}$ suit asymptotiquement une loi normale $N(0,1)$. Sinon, les valeurs p doivent être calculées empiriquement par permutation.

L'inconvénient de cette approche est que : 1) elle est limitée aux traits binaires, 2) elle ne permet pas d'inclure des covariables et 3) elle nécessite des permutations pour évaluer la signification du test statistique (calcul intensive).

Lien entre C-alpha et SKAT

Comme cela a été montré dans Wu et al. [84], C-alpha est un cas particulier de SKAT. En effet, notons par $T_C^1 = \sum_{j=1}^{k}(m_j - n_j p_0)^2$ la première partie de la quantité T_C. Puisque $\sum_{j=1}^{k} n_j p_0(1-p_0)$ est la moyenne de T_C sous l'hypothèse nulle, T_C^1 est le test C-alpha sans soustraire la moyenne. On peut voir aussi que $m_j = Y'X_j$ et $n_j = J'X_j$ avec $J = (1, \dots, 1)'$. On peut donc écrire le test C-alpha sous la forme suivante:

$$T_C^1 = (Y - p_0 J)'XX'(Y - p_0 J)' = (Y - \bar{Y})'XX'(Y - \bar{Y})'$$

C'est la version non pondérée de SKAT sans covariables.

En conclusion, les différentes approches décrites précédemment sont relativement équivalentes. Cependant, les approches KMR sont plus avantageuses à cause de la flexibilité offerte par la fonction de noyau. De plus, cette approche permet de conduire le test d'association en ajustant pour des covariables et tenir compte de la stratification de population. Les approches KMR nécessitent, cependant, de spécifier la fonction du noyau. Plusieurs choix sont possibles et ce choix n'est pas sans conséquence sur la performance de la méthode. A ce jour, il n'y a pas d'évidence sur le meilleur choix à faire et ce choix dépend du modèle génétique. Cependant, Wu et al, [77] suggèrent l'utilisation de noyaux linéaires si on suppose l'absence d'effets d'interaction entre les variants, et le noyau IBS dans le cas contraire. L'effet d'interaction peut être testé aussi par l'utilisation de noyaux quadratiques [84]. A notre connaissance, la comparaison des performances de ces deux fonctions de noyau n'a pas été faite.

Les tests d'association SNP-Set ont été comparé avec les tests de régression multivariée et le test simple-marqueur [75,76,77]. Pour le test simple-marqueur, la signification de l'unité génétique testée (gène) était celle du SNP ayant la meilleure signification corrigée par le nombre de tests (SNPs) de l'unité. Dans ces données simulées, les résultats montrent que le SNP-Set test a la meilleure puissance. Ces résultats sont encourageants et montrent que

l’analyse jointe de plusieurs marqueurs apporte une information meilleure sur la variabilité génétique que le test simple-marqueur, et que ce gain d’information peut se traduire par un gain de puissance. Ces conclusions ont été, cependant, obtenues dans des données simulées et sous certains modèles génétiques et caractéristiques génétiques qui ne sont peut-être pas représentatives de la variabilité du génome observée dans des données réelles. De plus, à notre connaissance, et à ce jour, aucune étude n’a comparé les tests SNP-Set et haplotypique dans un même jeu de données. Ici, nous nous sommes intéressés à cette problématique dans nos données réelles GWAS de la MP.

3.2.2 Evaluation dans les données génotypiques françaises de la MP

Matériel et méthodes

Données: L'étude est basée sur la cohorte française du GWAS de la MP. Elle contient 1039 cas et 1984 témoins génotypés pour 492929 SNPs (postQC).

Annotation des gènes: Nous avons défini les gènes comme unité d'analyse. Pour mieux couvrir les gènes, nous y avons ajouté 5 kb en amont et en aval. Nous avons utilisé les gènes autosomaux de la base de données « Ensemble Gene 63 Database (GRCh37.p3) ». Nous avons étudié les gènes qui contiennent au moins deux SNPs. Le nombre de ces gènes est 31263.

Analyses statistiques d'association :

Tests d'association haplotypiques: Nous avons utilisé le modèle logistique d'association haplotypique que nous avons décrit dans le chapitre 1, section 1.2.2.2. Ce test est implémenté dans le logiciel PLINK [14]. La première étape de ce test est d'inférer les haplotypes par l'algorithme EM et estimer leurs fréquences. Les haplotypes de fréquences < 1% ont été exclus du test. La signification du test haplotypique est celle du test omnibus, c'est-à-dire, associé à la comparaison globale des fréquences des haplotypes chez les cas et les témoins.

Test d'association SNP-Set (TEST5, section précédente): Le modèle que nous avons utilisé est décrit dans la section 3.2.1. Le test utilisé [84] est : $Q = (Y - \hat{\mu})'XX'(Y - \hat{\mu}) \sim \sum_{i=1}^{n} \lambda_i \chi^2_{(1),i}$

avec $\lambda_i, i = 1, \dots, n$ sont les valeurs propres de la matrice $P_0^{1/2} XX' P_0^{1/2}$ avec $P_0 = V - V\tilde{X}'\left(\tilde{X}V\tilde{X}\right)^{-1}\tilde{X}'V$, $V = \hat{\sigma}^2 I$, I est la matrice identité $n \times n$, $\tilde{X} = [1 \ Z]$ et Z est la matrice de covariables.

La signification du test peut être approximée comme nous l'avons montré dans la page 55 (TEST5, section précédente). Une autre approximation est la méthode de Davies [87]. Le test TEST5 ainsi que ces deux méthodes d'approximation sont implémentées dans le package R « SKAT » (http://www.hsph.harvard.edu/research/skat/download/). Nous avons utilisé l'approximation de Davies pour calculer la signification.

Résultats

Le nombre total de gènes ayant au moins deux SNPs (PostQC) est de 31263. Le test SNP-Set a été réalisé pour tous ces gènes. En revanche, le test haplotypique n'a pas pu être obtenu pour tous les gènes à cause du trop grand nombre d'haplotypes et/ou d'une mauvaise inférence des phases allèliques au sein du gène. Au total, 30607 gènes ont pu être testés avec le test haplotypique.

Distributions « gene-wide » des tests : La distribution du nombre de tests significatifs à différents niveaux de signification théorique obtenus sous chacun de ces tests est donnée dans la table 3.6. Pour les 30607 gènes testés par ces deux tests, la figure 3.8 montre la relation des résultats statistiques (P) obtenus avec SNP-Set vs Haplotypique.

Globalement, les distributions observées ne s'écartent pas trop de la distribution attendue. Pour un même seuil (α) de signification théorique, les taux d'associations positives du test Haplotypique et ceux du test SNP-Set sont assez similaires. La corrélation des résultats statistiques (-log10*P*) est 0.68. L'analyse de régression des ces deux variables montrent qu'elles sont fortement associés : l'hypothèse nulle d'indépendance des deux variables est rejetée ($P < 10^{-16}$). Cependant, comme illustré dans les figures 3.9 et 3.10, on remarque des différences notables dans la localisation des meilleurs signaux d'association obtenus sous chacun des tests.

Table 3.6- Distributions « gene-wide » des tests d'association. N= Nombre de gènes testés. Nombre et proportion de résultats significatifs au seuil théorique (α).

α	SNP-Set	(N=31263)	Haplotypique	(N=30607)
	Nombre	Proportion	Nombre	Proportion
5x10-2	1845	5.90E-02	1726	5.64E-02
10-2	391	1.25E-02	394	1.29E-02
10-3	43	1.38E-03	43	1.40E-03
10-4	8	2.56E-04	4	1.31E-04
10-5	1	3.20E-05	2	6.53E-05
10-6	0	0	0	0
10-7	0	0	0	0
10-8	0	0	0	0

Figure 3.8- Correspondance des résultats statistiques (-log10*P*) du test « Haplotypique » (abscisse) et du test « SNP-Set » (ordonné).

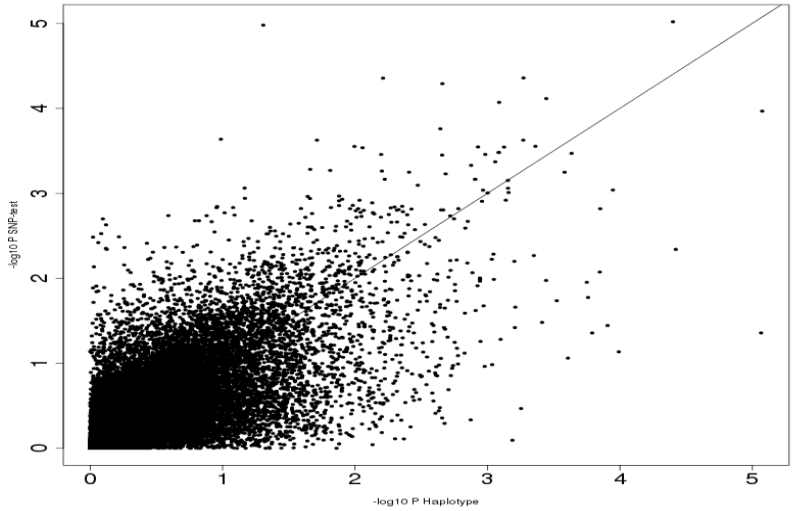


Figure 3.9- Manhattan-Plot: Test d'association haplotypique

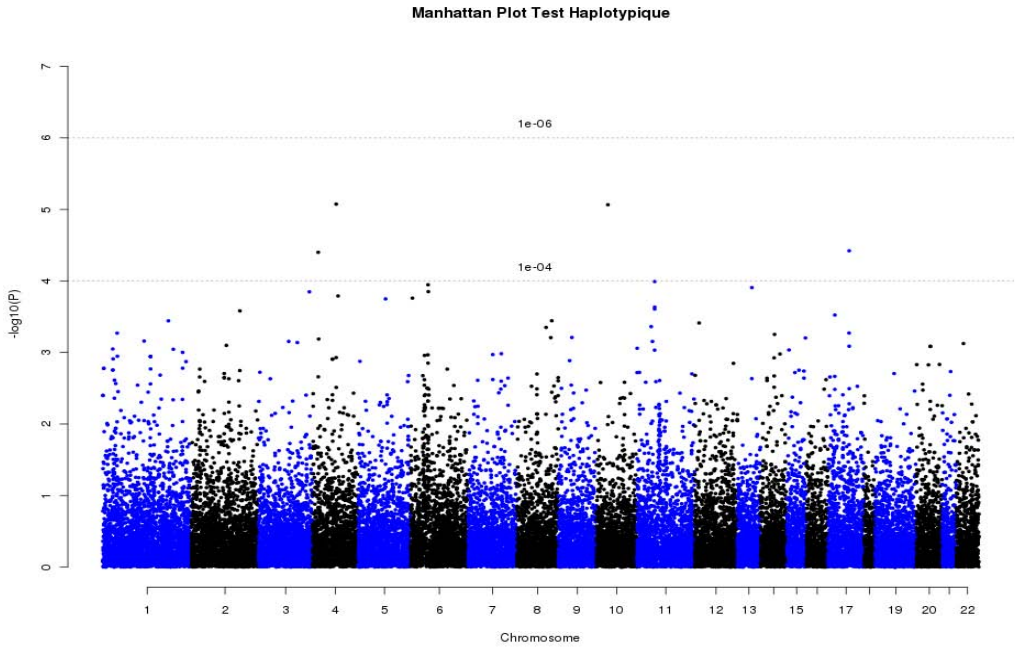
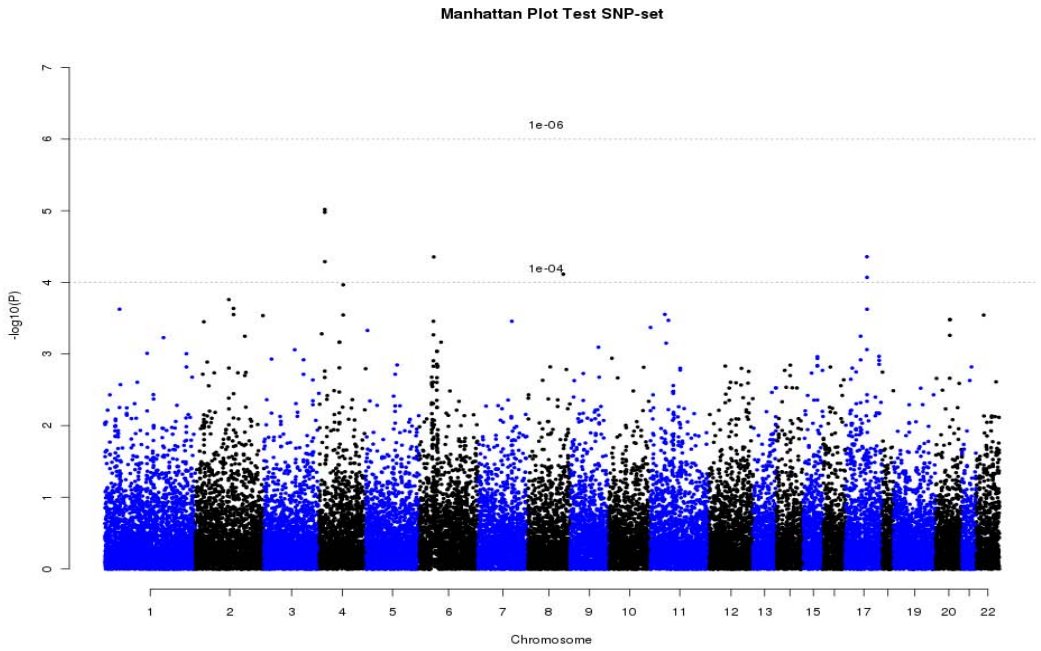
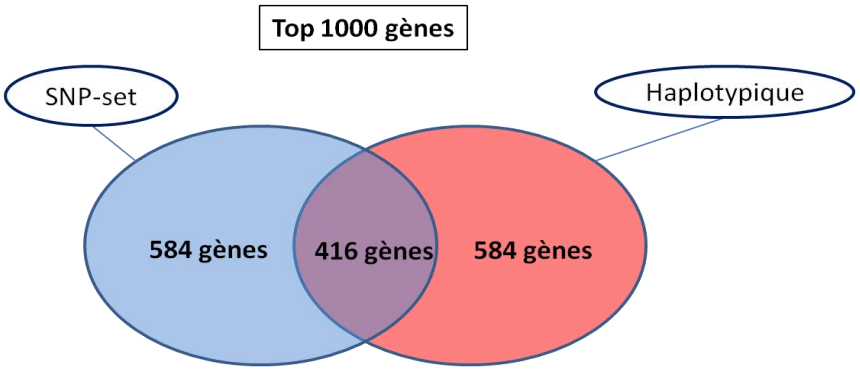


Figure 3.10- Manhattan-Plot: Test d'association SNP-Set



Concordance des signaux d'association : Pour chaque test, nous avons ordonné tous les signaux d'association par leur significativité statistique : de la plus petite valeur de P à la plus grande. Nous avons ensuite évalué la concordance des rangs des signaux (gènes) obtenus sous chaque test. Le diagramme suivant montre la concordance trouvée pour les 1000 gènes les plus associés sous l'un ou l'autre test. Près de la moitié des gènes (41.6%) sont identifiés parmi les 1000 meilleurs signaux d'association sous les 2 tests.



Cette concordance tombe à 23% et 11% lorsque l'on s'intéresse aux 100 et 10 gènes les plus associés, respectivement. Ainsi, au niveau des meilleurs signaux d'association, différents gènes sont identifiés par les tests.

Les résultats détaillés des gènes les plus associés (rang ≤10), sous l'un ou l'autre test, sont montrés dans la table 3.7. La table donne les caractéristiques principales des gènes: nombre de SNPs, longueur du gène et nombre de blocs de LD dans le gène. Deux gènes seulement sont identifiés par les deux tests : le premier est un gène connu de la MP (SNCA) alors que le deuxième (RP11-115L11.1) est adjacent à un autre gène connu de la MP (BST1). La table montre qu'un certain nombre des gènes les plus associés sous « Gene-Set » n'ont pas un petit nombre (>10) de SNPs (i.e., AL662797.2, BST1, K1AA1267). L'évidence d'association peut être plus forte sous le test Haplotypique que sous le test « Gene-Set », en dépit du grand nombre de *df* (i.e., SNCA, RP11-342D11.3).

En conclusion, ces premières analyses n'illustrent pas que la pénalité du *df* ai un impact majeur sur la performance du test haplotypique vis-à-vis du test « SNP-Set » à 1 *df*. Par ailleurs, nos analyses haplotypiques ne sont probablement pas optimales ; elles ont été conduites sans tenir compte des blocs de LD au sein du gène. Le test haplotypique au sein de ces blocs est certainement une stratégie d'analyse plus pertinente.

Table 3.7- Caractéristiques des 10 gènes les plus associés sous le test haplotypique ou « SNP-Set ». Les noms de gènes en gras sont des gènes connus de susceptibilité ou proches d'un de ces gènes. Les signaux d'association dont les rangs sont ≤10 sous les deux tests sont ombrés. Pour chaque gène, la valeur de *P* la plus petite est soulignée ; le signal d'association le plus significatif obtenu par chacun des tests est en rouge.

					Haplotypique			SNP-Set	
Chr	**Gène**	**(kb)**	**#SNP**	**# Bloc**	**df**	**P**	**Rang**	**P**	**Rang**
2	*IL1F8*	30.78	5	1	7	2.3E-03	97	1.7E-04	9
2	*AC016725.4*	183.25	13	2	10	1.0E-01	3420	2.3E-04	10
3	*RP11-95L3.2*	0.93	3	1	2	1.4E-04	9	8.4E-03	328
4	***BST1***	35.36	19	5	12	4.9E-02	1698	1.0E-05	2
4	***RP11-115L11.1***	0.67	3	1	3	4.0E-05	4	9.6E-06	1
4	***RP11-442P12.2***	0.38	7	1	6	2.2E-03	92	5.1E-05	5
4	***SNCA***	114.22	17	2	10	8.4E-06	1	1.1E-04	8
4	*RP11-167N19.2*	12.65	5	1	5	1.6E-04	10	4.4E-02	1619
6	***AL662797.2***	25.43	37	3	10	6.2E-03	252	4.4E-05	4
6	*MIR1236*	0.10	5	1	4	1.1E-04	6	9.2E-04	40
6	*PRRT1*	6.01	5	2	6	1.4E-04	8	1.5E-03	67
8	***7SK.235***	0.29	4	1	4	3.6E-04	17	7.7E-05	6
10	***RP11-342D11.3***	20.42	8	2	7	8.6E-06	2	4.4E-02	1614
11	*CTD-2119L1.1*	1.55	2	1	2	1.0E-04	5	7.3E-02	2585
13	*ATXN8OS*	24.33	8	2	7	1.2E-04	7	3.6E-02	1335
17	***AC217771.1***	2.18	8	2	6	5.4E-04	22	4.4E-05	3
17	***STH***	0.44	2	1	2	3.8E-05	3	4.6E-03	175
17	***KIAA1267***	162.83	12	1	4	8.2E-04	35	8.5E-05	7

Analyses haplotypiques des blocs de LD

Nous avons ré-analysé nos données en estimant les blocs de LD au sein des gènes. Cette estimation a été faite avec le logiciel PLINK [14], en utilisant la commande suivante :

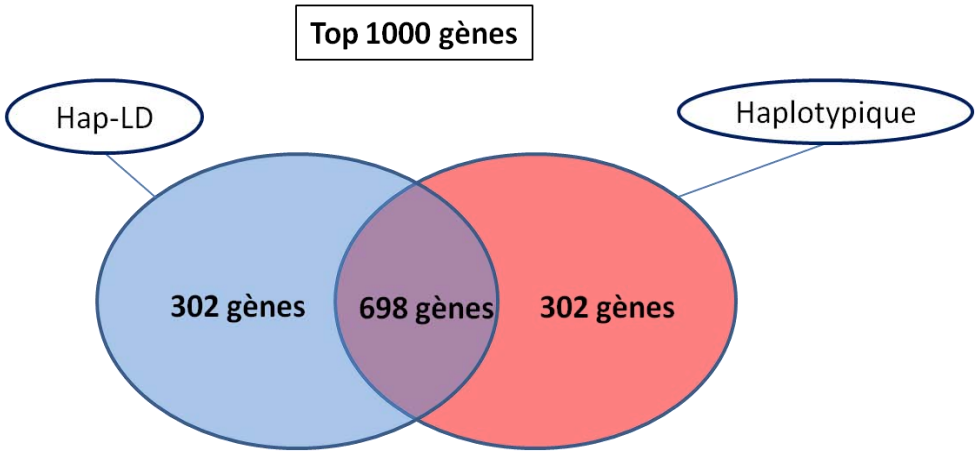
plink --file CHR -- blocks -- out out_blocks

L'unité d'analyse est donc maintenant le bloc de LD. Le test haplotypique du bloc de LD a été réalisé avec PLINK, sous le même modèle logistique que celui utilisé pour l'analyse haplotypique du gène entier. Au total, 77033 blocs de LD on été testés. Ces blocs appartiennent à 28884 gènes. Pour mesurer la signification au niveau du gène, nous avons pris la valeur P du bloc le plus associé et l'avons multipliée par le nombre de blocs testés dans le gène. La table 3.8 montre les distributions du test haplotypique tenant compte du LD lorsque l'unité d'analyse est le bloc (« Hap-bloc», P non corrigée) ou le gène (Hap-LD). Le diagramme suivant montre la concordance des résultats pour les 1000 signaux les plus associés.

Table 3.8- Test haplotypique tenant compte du LD : Distribution des statistiques du test de l'association au niveau du bloc (Hap-bloc) et au niveau du gène (Hap-LD)

Nombre et proportion de résultats significatifs au seuil nominal (a). N= Nombre de tests

α	Hap-bloc	N=77033	Hap-LD	N=28884
	Nombre	Proportion	Nombre	Proportion
5x10-2	4311	5.60E-02	1529	5.29E-02
10-2	925	1.20E-02	372	1.29E-02
10-3	97	1.26E-03	40	1.38E-03
10-4	14	1.82E-04	8	2.77E-04
10-5	5	6.49E-05	5	1.73E-04
10-6	3	3.89E-05	1	3.46E-05
10-7	1	1.30E-05	1	3.46E-05
10-8	0	0	0	0



Il est clair que le test haplotypique basé sur les blocs du LD donne de meilleurs niveaux de signification, même après correction pour le nombre de blocs testés dans le gène.

Par ailleurs, la concordance des résultats obtenus par cette approche et celle qui ignore les blocs de LD est assez forte : près de 70% des gènes sont trouvés par les deux approches parmi leurs 1000 meilleurs signaux d'association respectifs. En considérant les 100 ou 10 meilleurs signaux d'association, la concordance reste importante, 53% et 43%, respectivement.

Discussion et conclusion

Quel sont les facteurs qui peuvent influencer les différences de distributions de ces tests ?

Nous avons étudié deux facteurs principaux : le nombre de SNPs et le pattern de déséquilibre de liaison dans le gène. Le pattern de LD a été mesuré comme la moyenne du r^2 de toutes les combinaisons possibles des paires de SNPs dans le gène.

Nous avons classé les gènes dans quatre classes selon les quartiles de la distribution des valeurs du facteur étudié : -- 1ère classe = [0- 1er Quartile (Q1) [, -- 2ème classe = [Q1-Médiane[, -- 3ème classe = [Médiane-3ème Quartile (Q3) [et 4ème classe = [Q3-Max]. Dans chaque classe, nous avons comparé la proportion de gènes dont la signification est plus petite qu'une valeur nominal donnée (α= 0.05, 0.01 et 0.001).

1) **Nombre de SNPs dans le gène** (Figure 3.11): les valeurs respectives du Q1, de la médiane et du Q3 de la distribution du nombre de SNPs dans les gènes sont trois, cinq et neuf.

   Nous observons que le nombre des résultats significatifs du test « SNP-Set » augmente dans la classe des gènes ayant le plus grand nombre de SNPs.

   Cette relation n'est pas retrouvée pour le test « Hap-LD ». Une explication possible est que les gènes ayant le plus grand nombre de SNPs sont des gènes longs. Comme le montre la figure 3.12, le nombre de blocs de LD est positivement corrélé au nombre de SNPs dans le gène. Ainsi, les gènes ayant un grand nombre de SNPs ont aussi un grand nombre de blocs de LD et leur signification statistique avec le test « Hap-LD » est donc particulièrement pénalisée par la correction de Bonferroni.

Figure 3.11- Proportions de gènes significatifs à différents seuils nominaux de signification par classe de gènes selon le nombre de SNP.

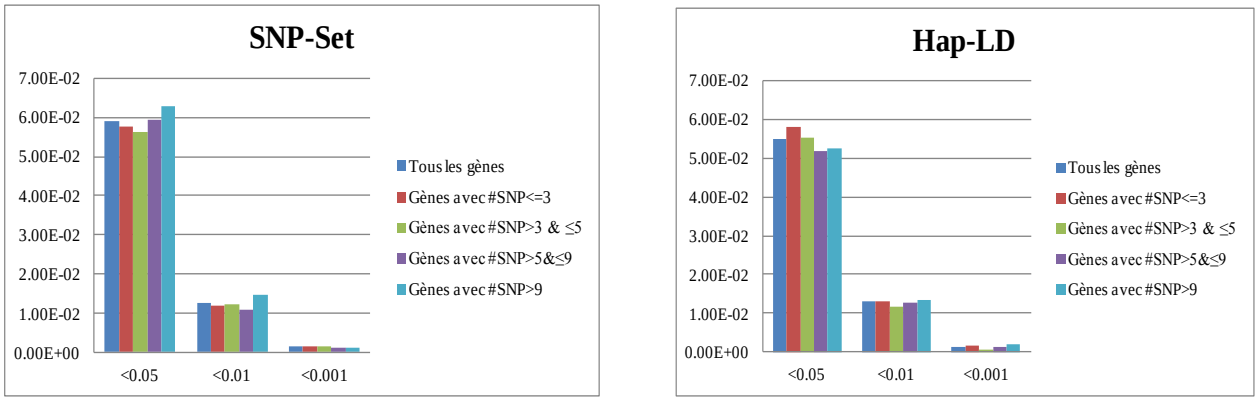
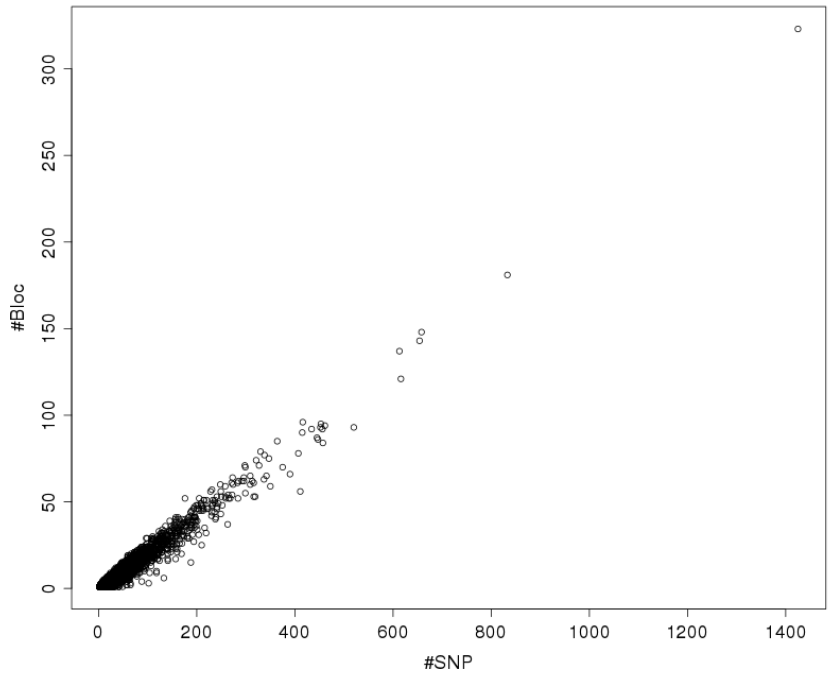
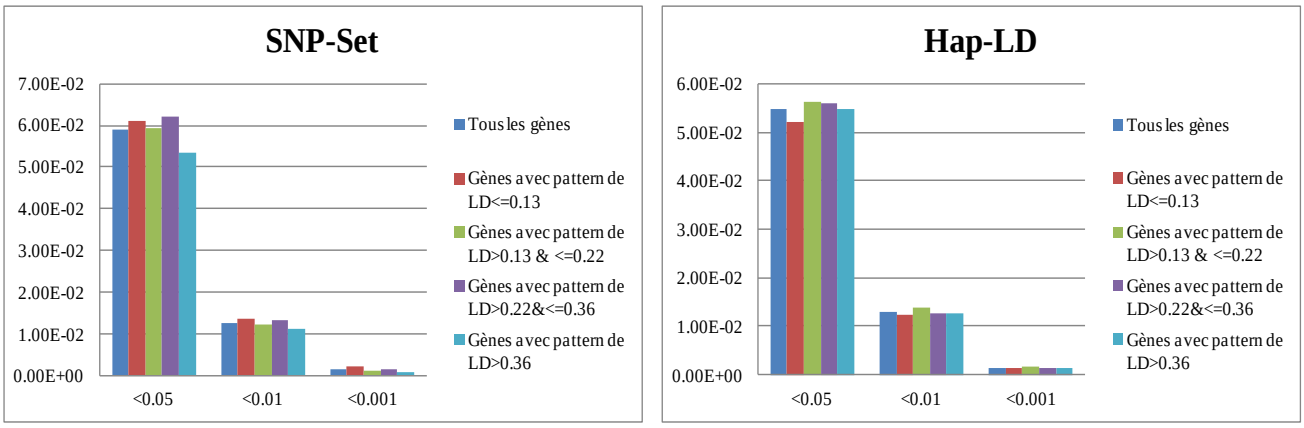


Figure 3.12- Relation entre le nombre de SNPs et le nombre de bloc de LD dans le gène.



2) **Pattern de LD dans le gène** (Figure 3.13): les valeurs respectives du Q1, de la médiane et du Q3 de la distribution du pattern de LD dans les gènes sont 0.13, 0.22 et 0.36. Nous observons que la proportion de résultats significatifs est plus petite dans la 4ème classe (i.e. gènes ayant de forts niveaux de LD) avec le test « SNP-Set » mais pas avec le test « Hap-LD ».

Figure 3.13- Proportions de gènes significatifs à différents seuils nominaux de signification par classe de gènes selon le pattern de LD.



Globalement, ces analyses ne montrent pas de relation importante entre le nombre de SNPs ou le pattern de LD dans le gène et la signification statistique des tests « SNP-Set » et « Hap-LD ».

Qu'avons-nous gagné par rapport au test simple-marqueur?

La figure 3.14 montre la variation des significations statistiques du test Hap-LD et du SNP-Set avec celles du test simple-marqueur (P corrigée par le nombre de SNPs du gène). On remarque que les niveaux de signification du test SNP-Set et ceux du test simple-marqueur sont assez similaires. En revanche, l'évidence d'association pour un certain nombre de gènes est plus forte sous le test haplotypique que sous le test simple-marqueur, où dans certain cas l'association n'est pas significative statistiquement. La figure 3.15 détaille ces variations pour les 1000 meilleurs signaux d'association. Elle montre le gain pouvant être apporté par le test Hap-LD par rapport au test de simple-marqueur, dans le groupe des 1000 meilleurs signaux d'association.

Figure 3.14- Relation entre les niveaux de signification des tests « Hap-LD » et « simple marqueur » (droite) et ceux des tests « SNP-Set » et « simple marqueur » (gauche).

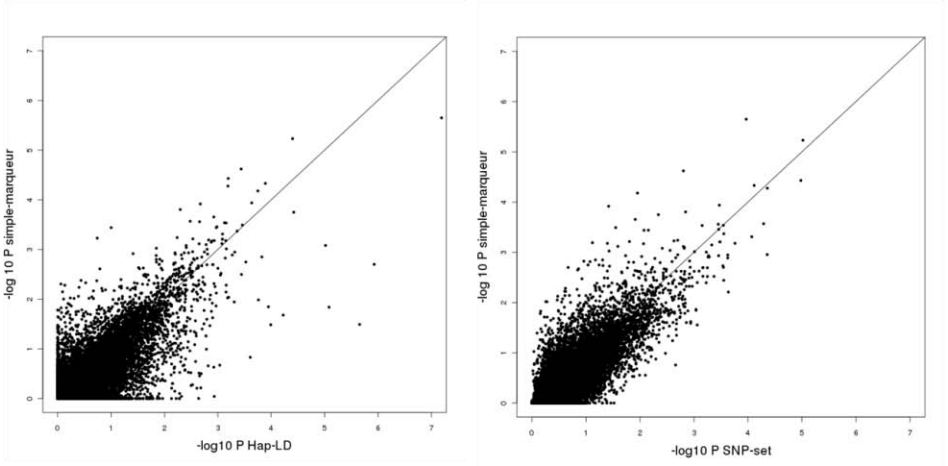
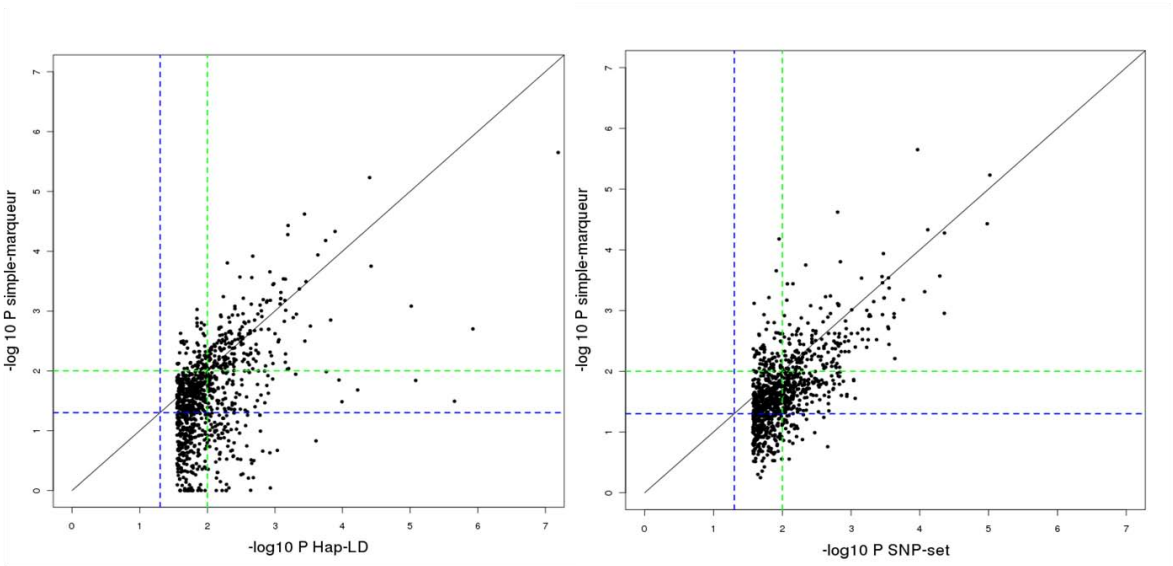


Figure 3.15- Niveaux de signification des 1000 gènes les plus associés de l'approche Hap-LD (gauche) et de l'approche SNP-Set (droite) vis-à-vis de l'approche simple-marqueur. La ligne bleue montre la signification à 5%. La ligne verte montre la signification à 1%.



Enfin, nous présentons dans la table 3.9 les niveaux de signification (P) des dix meilleurs gènes détectés par l'un des trois tests multi-marqueur. Elle montre aussi le niveau de signification du test simple-marqueur, après correction pour le nombre de SNPs du gène. Plusieurs des gènes montrés dans cette table sont des gènes connus de la MP: SNCA, BST1 et GAK. Deux gènes (SNCA et BST1) ont été préalablement identifiés par nos analyses simple-marqueur GWAS. Pour SNCA, l'évidence d'association reste forte ($P< 1.1\times10^{-4}$) sous tous les tests multi-marqueur : elle est la plus forte avec le test haplotypique « Hap-LD » ($P=6.5\times10^{-8}$). Pour le gène BST1, les deux meilleurs niveaux de signification sont obtenus avec SNP-Set et aussi avec le test simple-marqueur. Il est intéressant de noter, pour le gène GAK, l'évidence d'association forte du test haplotypique « Hap-LD » ($P= 6\times10^{-5}$), malgré le grand nombre de blocs (N=17) dans ce gène et le *df* (4) du test. De plus, cette évidence d'association est bien meilleure que celle obtenue avec le test simple-marqueur ($P=2.5\times10^{-3}$) ou le test SNP-Set. Globalement, par rapport aux résultats du test simple-marqueur, on observe plus de signaux d'association nouveaux avec le test Hap-LD qu'avec le test SNP-Set. Nous ne savons pas si ces signaux d'association sont des faux positifs ou non. Pour les 24 loci connus de la MP, sept, sept et cinq gènes ont des rangs inférieurs ou égaux à 10 sous les tests « SNP-Set », « Hap-LD » et « Haplotypique » respectivement. L'extension de cette étude à une étape de réplication serait intéressante. Ceci pourrait se faire au sein de nos collaborations de l'IPDGC.

Table 3.9- Résultats détaillés pour les 10 meilleurs signaux d'association de l'un des trois tests multi-marqueur (Haplotypique, Hap-LD et SNP-Set) et P du test simple-marqueur. Les noms de gènes en gras sont des gènes connus de susceptibilité ou proches d'un de ces gènes. Les signaux d'association dont les rangs sont ≤10 sous les deux tests sont ombrés. Pour chaque gène, la valeur de *P* la plus petite est soulignée ; le signal d'association le plus significatif obtenu par chacun des tests est en rouge.

			Haplotypique		Hap-LD			SNP-Set		Simple-marqueur
Chr	**Gène**	**(kb)**	**df**	**P**	**#bloc**	**df**	**%P**	**P**	**#SNP**	***P**
2	*IL1F8*	30.78	7	2.3E-03	1	3	1.2E-03	1.7E-04	5	6.6E-04
2	*AC016725.4*	183.25	10	1.0E-01	2	4	4.9E-03	2.3E-04	13	6.2E-03
3	*RP11-95L3.2*	0.93	2	1.4E-04	1	2	1.0E-01	8.4E-03	3	3.6E-04
4	***GAK***	83.10	13	1.8E-02	17	4	6.0E-05	5.0E-03	2	2.5E-03
4	***BST1***	35.36	12	4.9E-02	5	2	6.4E-04	1.0E-05	19	3.7E-05
4	***RP11-115L11.1***	0.67	3	4.0E-05	1	3	4.0E-05	9.6E-06	3	5.9E-06
4	***RP11-442P12.2***	0.38	6	2.2E-03	1	5	3.3E-03	5.1E-05	7	2.7E-04
4	***RP11-115D19.1***	175.15	14	1.2E-03	22	2	1.2E-06	2.8E-04	6	5.4E-04
4	***SNCA***	114.22	10	8.4E-06	2	2	6.5E-08	1.1E-04	17	2.2E-06
4	*RP11-167N19.2*	12.65	5	1.6E-04	1	2	9.5E-03	4.4E-02	5	6.8E-02
6	***AL662797.2***	25.43	10	6.2E-03	3	8	3.0E-03	4.4E-05	37	1.1E-03
6	***MIR1236***	0.10	4	1.1E-04	5	4	1.1E-04	9.2E-04	1	2.9E-03
6	***PRRT1***	6.01	6	1.4E-04	2	3	9.6E-06	1.5E-03	5	8.3E-04
7	*AC019117.1*	91.80	14	2.3E-01	23	2	8.3E-06	1.6E-02	6	3.8E-03
8	***7SK.235***	0.29	4	3.6E-04	1	3	1.3E-04	7.7E-05	4	4.7E-05
10	***RP11-342D11.3***	20.42	7	8.6E-06	2	2	2.6E-01	4.4E-02	8	6.5E-02
11	*CTD-2119L1.1*	1.55	2	1.0E-04	1	2	1.0E-04	7.3E-02	2	3.3E-02
11	*DYNC2H1*	370.43	15	5.4E-01	62	5	2.2E-06	4.8E-01	8	4.2E-03
13	*ATXN8OS*	24.33	7	1.2E-04	8	2	2.4E-01	3.6E-02	2	2.3E-02
17	***AC217771.1***	2.18	6	5.4E-04	2	2	6.5E-04	4.4E-05	8	5.3E-05
17	***STH***	0.44	2	3.8E-05	1	2	3.8E-05	4.6E-03	2	1.8E-04
17	***KIAA1267***	162.83	4	8.2E-04	1	4	8.2E-04	8.5E-05	12	4.9E-04

%Corrigée pour le nombre de blocs; *Corrigée pour le nombre de SNPs

Tests multi-marqueur dans les données d'imputation

Toutes nos comparaisons ont été effectuées sur des données génotypées. Toutefois, elles pourraient être effectuées sur des données imputées. Dans ce cas, nous n'avons pas le génotype exact du SNP imputé mais plutôt l'estimation du nombre d'allèle mineur représentée par la dose allélique (Section 3.1.2, méthodes d'imputation). L'application du test « SNP-Set » sur les données imputées ne pose pas de problème et elle est facile à faire. En revanche, le test haplotypique n'est pas faisable puisqu'on ne peut pas estimer les haplotypes en se basant sur les doses alléliques. Une alternative serait d'utiliser le génotype le plus probable au lieu de la dose allélique mais ceci est déconseillée par la littérature [88] à cause de l'imprécision de l'estimation des haplotypes.

En conclusion, notre étude comparative des tests multi-marqueur dans les données GWAS réelles de la MP, ne suggère pas que le test SNP-Set soit globalement plus puissant que le test simple-marqueur, comme cela a été rapporté dans des données simulées. De même, nous n'observons pas une meilleure performance du test « SNP-Set » par rapport à celle du test haplotypique. Au contraire, dans nos données, il semble que le test haplotypique, en particulier lorsque l'on tient compte du LD, permette de mieux capturer la variabilité génétique locale que le test SNP-Set.

3.2.3 Maladie Commune – Variant rare

Il est bien connu que la puissance du test simple-marqueur est faible pour détecter l'association avec des variants rares. Pour analyser les données de séquençage dans le but de détecter de tels variants, des nouvelles méthodes statistiques ont été développées. Dans cette partie, nous exposons les travaux abordant cette problématique.

Les variants rares peuvent-ils contribuer à l'étiologie des maladies complexes?

L'hypothèse des études d'association pour les variants rares est: "Trait Commun - Variants Rares Multiples" (abréviation en anglais: CDMRV) [89]. Cette hypothèse suppose que plusieurs variants rares contribuent collectivement à la susceptibilité des traits complexes. Des études récentes basées principalement sur du séquençage ont identifié des variants rares multiples impliqués dans la susceptibilité de traits complexes. Un exemple remarquable est le cas du cancer du sein. L'association avec dix gènes qui contiennent plusieurs mutations rares a été identifiée [90]. Il a été aussi montré que certains variants rares dans les gènes SLC12A3, SLC12A1 et KCNJ1 contribuent à la réduction de la pression sanguine et à la protection de l'hypertension [91].

Nouvelles méthodes alternatives au test simple-marqueur: méthodes « Collapsing »

Des méthodes alternatives au test simple-marqueur ont été proposées. Ces méthodes ne consistent plus à tester l'effet individuel du variant mais plutôt l'effet cumulatif des variants rares d'une même unité génomique, par exemple le gène.

Le principe des tests d'association qui regroupe l'information de plusieurs variants rares a été proposé pour la première fois en 2007 par Morgenthaler, S. et Thilly, W. G. [92]. Le test proposé, « Cohort Allelic Sums Test (CAST)», consistait à comparer le nombre de malades et de témoins qui portent au moins un allèle rare à tous les variants d'un gène, en utilisant un test de χ^2 standard ou le test de Fisher exact.

En 2008, Li et Leal [93] ont montré, à travers des simulations, que la puissance de CAST est plus importante que celle du test simple-marqueur. Ils ont aussi évalués le test de la régression multivariée, en incluant tous les variants rares dans le modèle.

Une autre hypothèse est celle de la contribution jointe de variants rares et communs à la susceptibilité de la maladie [94,95]. Li et Leal ont proposé une méthode appelée « Combined Collapsing Multivariate (CMC) » qui consiste à regrouper les variants rares dans une seule

méga-variable et ensuite la tester avec les variants communs dans un modèle multivarié (i.e régression multiple, test de Hotelling).

Les principales conclusions de cette étude sont:

1- Le test simple-marqueur est le test le moins puissant à cause de deux problèmes majeures : 1) le problème de test multiple et 2) la faible puissance pour détecter les variants rares. De plus, pour des variants très rares, le test de χ^2 n'est plus valide et il faut donc utiliser le test de Fisher exact. Celui-ci est conservateur.
2- Le test multivarié est plus puissant que le test simple-marqueur malgré le problème du nombre de degrés de liberté.
3- Le test CAST est plus puissant que le test multivarié. Cela semble normal parce que tous les variants rares sont regroupés dans une variable unique et donc le nombre de degré de liberté est réduit à un. En revanche, la puissance de CAST diminue dramatiquement lors de l'inclusion de variants non causaux, particulièrement s'ils ont des fréquences relativement élevées (MAF=0.1).
4- Le test CMC est plus robuste à l'inclusion de variants non causaux et présente la meilleure puissance dans ces cas là.

Dans l'étude de Li et al., tous les tests statistiques sont proposés pour des données de type cas-témoins. En plus, les tests utilisés ne permettent pas d'inclure des covariables, ce qui empêche de contrôler les effets de facteurs de confusion comme la stratification de population.

Ces tests peuvent, cependant, être généralisés pour les traits quantitatifs en utilisant des modèles de régression classique, et inclure des covariables dans le modèle.

Modèle de régression: Introduction des covariables et analyse de traits binaires

En 2010, Morris et Zeggini [96] ont proposé une extension de CAST en utilisant le modèle de régression. Ce modèle se base sur le nombre de SNP dans le gène où un individu porte au moins une copie de l'allèle rare. Cette étude rapporte que la puissance de cette approche est supérieure ou égale à la puissance de CAST. En effet, CAST peut présenter une perte de puissance lorsque la région génomique testée contient un grand nombre de variants rares avec des fréquences alléliques plus ou moins grandes.

Quel seuil de MAF doit-t-on choisir pour définir les variants rares?

En général, un variant rare est défini par sa fréquence plus petite que 1% [97]. Puisque le modèle génétique des traits complexes est souvent inconnu, deux seuils de MAFs, 1% et 5%, sont souvent utilisés pour filtrer les variants qui rentrent dans le modèle de "collapsing".

Le choix du seuil est crucial. Et si on n'a pas choisi le bon seuil?

Le modèle génétique des maladies complexes est inconnu. On ne connait pas les valeurs des MAFs des SNPs impliqués dans la maladie. Pour éviter le choix d'un seuil de sélection de variants, deux approches ont été proposées.

1- Pas de filtre de SNPs mais plutôt des poids

Madsen et Browning [98] ont proposé, en 2009, une méthode pour des traits binaires qui permet d'inclure tous les variants (fréquents et rares) dans le modèle de collapsing, en donnant plus de poids aux SNPs rares et moins de poids aux SNPs fréquents. La fonction de poids proposée est l'inverse de la variance de chaque variant estimé dans le groupe des témoins: $1/\sqrt{N_u MAF_u(1-MAF_u)}$ avec N_u et MAF_u sont respectivement le nombre des témoins et les fréquences alléliques estimées chez les témoins.

L'utilisation de cette fonction de poids revient pratiquement à exclure les variants ayant des MAFs >0.01. Le test d'association proposé est un test non paramétrique de rang (wilcoxon) et la signification est évaluée empiriquement par permutation des phénotypes parce que les MAF sont calculés dans les données dans le groupe de témoins.

2- Pas de filtre, pas de poids mais des seuils multiples

Price et al [99] ont proposé une généralisation du test de Madsen et Browning pour des traits quantitatifs. Ils ont proposé aussi un nouveau test d'association basé sur le choix de plusieurs seuils de MAF (VT). Le principe est le suivant:

1- Définir toutes les *N* valeurs de MAF (i.e S_1, ..., S_N) dans les données comme seuil de MAF pour la sélection de variants;

2- Créer les *N* variables de collapsing qui regroupent les variants satisfaisant la condition MAF < S_i, $i=1,...,N$;

3- Faire le test d'association *N* fois via des modèles de régression classique et évaluer la signification empiriquement par permutation des phénotypes.

Performance des méthodes: données génétiques simulées ou données de séquençage?

Les données idéales pour appliquer les méthodes de « collapsing » sont les données de séquençage. A ce jour, il n'y a pas d'étude de maladies complexes basées sur le séquençage tout génome.

Les études précédentes ont simulé des données génétiques et phénotypiques pour évaluer la performance des différentes méthodes citées. Cela n'a jamais été fait dans des vraies données de séquence tout génome jusqu'en octobre 2010 par le groupe GAW17 « Genetic Analysis Workshop 17 ».

- Propriétés statistiques des tests d'association de type « collapsing »

Nous avons participé à cet atelier dans le but d'évaluer les méthodes statistiques pour la détection d'association de variants rares dans le cadre de traits complexes. Les données génétiques fournies sont issues des séquences réelles du projet de 1000Genomes. A partir de ces sujets, des familles ont été construites et les transmissions génétiques au sein des familles ont été générées selon les lois de Mendel et distances génétiques entre variant. Pour chacun des individus, échantillon en population et familial, plusieurs phénotypes (traits qualitatifs et quantitatifs) ont été simulés.

Données génétiques: Les données génétiques fournies sont issues du projet 1000genomes. Elles incluent une partie (i.e. 24487 SNPs autosomaux) de l'exome. Ces SNPs appartiennent à 3205 gènes. Leurs MAFs varient entre 0.000717 (privé) et 0.5. Parmi ces SNPs, 9433 (38.4%) sont des variants privés. Les données génotypiques ont été fournies dans deux échantillons : (1) 697 sujets non apparentés et (2) huit grandes familles avec au total 697 sujets apparentés :

1. Echantillon de sujets non-apparentés: les 697 sujets non apparentés sont issus des quatre populations de HapMap : 156 Européens (CEU + Tuscan), 216 Chinois (Denver et Han), 105 Japonais, 220 Africains (Luhya et Yoruba).
2. Echantillon de huit familles (697 sujets apparentés): Dans ces familles, les couples fondateurs et les conjoints (N=202) ont été aléatoirement sélectionnés parmi les sujets du premier échantillon.

Phénotypes simulés : Quatre traits complexes ont été simulés. Un trait binaire (maladie) avec une prévalence de 30% et trois traits quantitatifs Q1, Q2, Q4. Les simulations ont été répétées jusqu'à obtenir 200 réplicas. Durant la simulation des traits, les données génotypiques sont restées fixes.

Pour désigner les variants impliqués dans les traits, des informations biologiques à priori des pathways ont été utilisées. Le pathway VEGF de la base de données Kyoto Encyclopedia of Genes and Genomes (KEGG) a été principalement utilisé pour désigner les gènes qui influencent les traits.

Dans notre étude, nous nous sommes intéressés au trait quantitatif Q1. Ce trait est influencé par 39 SNPs (rares et fréquents) dans 9 gènes. Le nombre de variants causaux par gène varie de 1 à 11 ; leurs MAFs sont comprises entre 0.07% et 16.5%. Tous ces variants sont à risque, c.à.d. l'allèle mineur est associé avec l'augmentation de Q1.

Analyses statistiques : Nous avons comparé les différentes variantes des méthodes « collapsing » qui différent par la façon de modéliser l'information génétique. Notre logique de comparaison a été la suivante :

1- Comparer les tests qui filtrent les variants aux tests qui ne filtrent pas ;
2- Comparer les tests qui utilisent tous les variants avec pondération à ceux qui utilisent le seuil multiple de MAF et aux tests CMC;
3- Comparer les tests qui utilisent le MAF = 5% comme filtre de SNPs aux tests qui utilisent le MAF=1%, dans les deux versions de « collapsing » : Absence/présence et Proportion.
4- Comparer les tests de « collapsing » au test simple-marqueur.
5- Comparer la performance de ces tests dans les données de population et dans les données familiales.

Dans les données de population, nous avons utilisé le modèle de régression linéaire. L'analyse des données familiales, a été conduite avec le test « Measure Genotype », basé sur un modèle linéaire mixte.

Estimation de la puissance et de l'erreur de type 1 : Pour estimer la puissance et l'erreur de type 1 des différentes approches, nous avons utilisé les 200 réplicas de Q1. La puissance a été évaluée pour chacun des neufs gènes causaux. Nous avons sélectionnés 7 gènes ne contenant

aucun variant causal et situés sur d'autres chromosomes que les gènes causaux. Ces gènes ont été sélectionnés aléatoirement parmi l'ensemble des gènes ayant des caractéristiques proches (nombre de SNPs et distributions des MAFs) de celles des gènes causaux. L'erreur de type 1 a été est évaluée pour chacun de ces sept gènes non-causaux. Ces deux quantités ont été estimées au seuil de signification α = 5% et elles mesurent la proportion de réplicas dans lesquels la valeur *P* du test d'association est plus petite ou égale à α.

Résultats

Données de population :

1. Les taux d'erreur de type 1 ne sont pas bien contrôlés. Ils varient par gène et par approche. Ils peuvent être plus, ou moins, élevés que les valeurs seuils théoriques.
2. Les approches qui utilisent les SNPs individuellement (simple-marqueur et CMC) ont tendance à être plus libérales que les autres. Il faut noter que plusieurs SNPs ont des fréquences alléliques spécifiques à chaque population (CEU, AFR, JPT et CHB). Or, nous avons observé que la moyenne de Q1 était différente entre les quatre populations. Le processus de simulation (génome fixé) ne permettait pas de s'affranchir de ces différences entre population. Ainsi, nous étions peut-être face à un problème de stratification de population. Nous avons donc effectué une analyse en composante principale et ajusté les modèles de la régression par rapport aux cinq premiers axes. Après cette étape, un faible nombre de tests montraient un taux de faux positifs plus grand qu'attendu (Avant correction: 20/70 taux de faux positifs > 0.05. Après correction : 3/70. (70 = 7 gènes × 10 approches)).
3. Nous avons observé que le choix du seuil de MAFs a beaucoup d'impact sur la puissance des approches Absence/présence et Proportion. Un résultat surprenant est que le choix du seuil à 5% montrait une puissance plus importante que celui de 1%, même lorsque les variants causaux ont tous des MAFs <1%. Ceci pourrait s'expliquer par l'existence de corrélations (i.e. LD) entre les SNPs causaux et non-causaux.
4. L'approche VT qui ne requiert pas un choix de seuil de MAF ne semble pas plus puissante que les approches qui utilise un seuil fixe de MAF.

Données familiales :

1. Les taux d'erreur de type 1 étaient mieux contrôlés par rapport aux données de population.

2. Le choix du seuil de MAF avait là encore un impact important. La puissance était meilleure avec la valeur seuil de 5% que 1%.
3. Trois gènes étaient détectés avec de bons niveaux de puissance (>80%), dont deux qui étaient aussi détectés dans les analyses en population. Le gène détecté spécifiquement dans les données familiales contient un seul variant causal et qui est rare (MAF<1%). Ainsi, un seul fondateur portant cet allèle rare a suffit pour l'introduire dans la famille, le rendant donc fréquent dans les familles. Pour les autres gènes, la puissance restait relativement petite.

En conclusion, dans notre étude, nous n'avons pas identifié d'approche qui soit globalement meilleure que les autres. Les performances relatives des méthodes variaient selon le gène testé. Pour la détection d'association de variants rares, la puissance de l'analyse d'association en données familiales peut s'avérer nettement plus élevée que celle obtenue en données de population. La performance du design familial est d'autant plus performante que les familles sont grandes. De plus, notre évaluation a été faite dans un échantillon de familles où la distribution du phénotype est aléatoire. On peut s'attendre à ce que la performance de l'analyse d'association dans des données familiales soit encore meilleure lorsque les familles sont sélectionnées sur des sujets ayant des valeurs extrêmes au trait étudié. L'évaluation de ces méthodes dans des données sélectionnées serait intéressante à développer.

Finalement, les approches de type « collapsing » ont une meilleure puissance par rapport au test simple-marqueur. En revanche, les niveaux de puissance que nous avons observés sont relativement petits, même au seuil de signification de 5%. Pour des critères de signification stricts (i.e. 1.6×10^{-6} = 0.05/30000 gènes), les niveaux de puissance seront évidement plus faibles. Pour avoir une bonne puissance (>80%), il faut donc des grandes tailles d'échantillons (plusieurs dizaines de milliers de sujets), même avec ces nouveaux tests de type « collapsing ».

3.2.4 Conclusions

A ce jour, le coût d'une analyse de maladies complexes par séquençage du génome ou de l'exome reste donc prohibitif. Un design alternatif a été récemment suggéré : séquencer une

partie des sujets d'une étude GWAS et imputer les sujets restants en se basant sur les données de séquence phasées du premier groupe.

Les données de « pseudo-séquençage » un design alternatif au séquençage ?

Il a été montré que cette stratégie est relativement puissante et permet d'apporter un gain de puissance assez important. Zawistowski et al, [100] ont comparé la puissance du test « collapsing » dans trois stratégies d'étude (Figure 3.16):

1) Analyse d'association de 200 sujets (tous complètement séquencés) ;
2) Analyse d'association de 2000 sujets (tous complètement séquencés) ;
3) Analyse d'association de 2000 sujets d'une étude GWAS (200 complètement séquencés et 1800 imputés).

Comme le montre la figure 3.16, le séquençage de 2000 sujets montre la meilleure puissance (66%). Cependant, le design alternatif, basé sur l'imputation de 1800 sujets de GWAS à partir de 200 sujets séquencés, montre une puissance meilleure (48%) que celle du design du séquençage d'un échantillon de petite taille.

Une autre étude récente [88] a comparé la puissance des test haplotypique et « collapsing » dans les données de génotypage et de pseudo-séquençage. La figure 3.17 montre que le test haplotypique est plus puissant que le test « collapsing » dans les données de génotypage. En revanche, les tendances s'inversent dans les données de pseudo-séquençage. La première conclusion conforte l'hypothèse de l'intérêt d'utiliser des tests d'association haplotypiques pour la détection de variants rares. La deuxième conclusion, comme l'explique les auteurs, est due à l'imprécision des estimations des haplotypes dans les données d'imputation.

Figure 3.16- Design alternatif au séquençage : Pseudo-séquençage

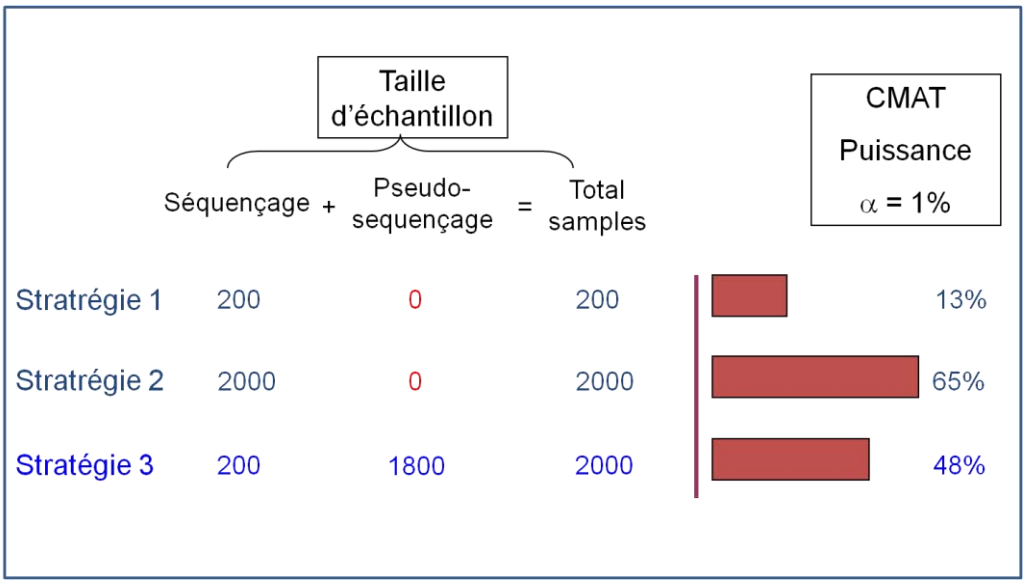
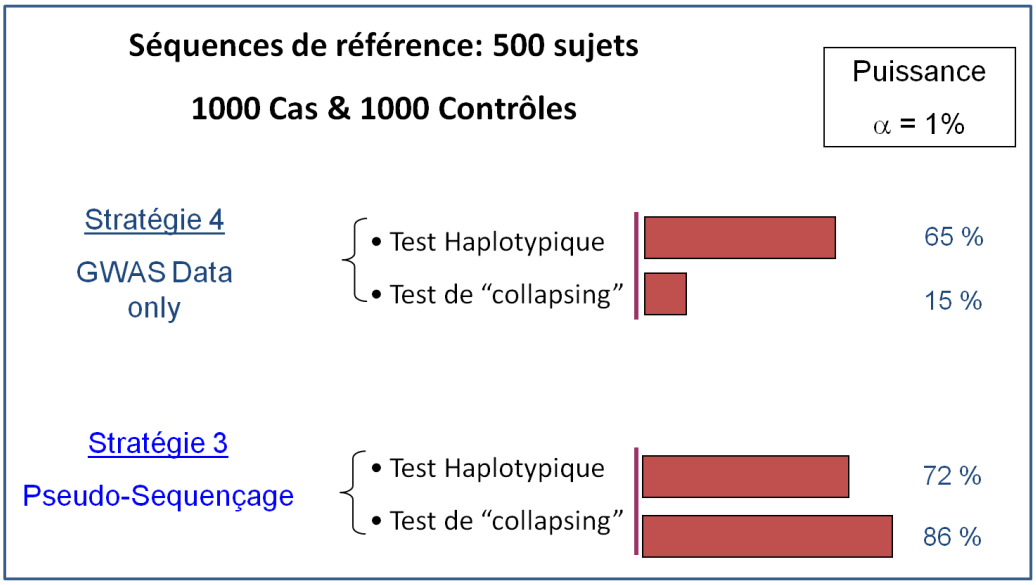


Figure 3.17- Puissance du test haplotypique et test « collapsing » dans les données de génotypage et dans les données de pseudo-séquençage



L'intérêt du design « pseudo-séquençage » pour la détection d'association de variants rares est très dépendant des données de séquence servant de bases aux imputations des données GWAS. Le nombre et caractéristiques des sujet séquencés sont deux facteurs majeurs. La probabilité de détecter l'association dans les données « pseudo-séquençage » dépend de la présence de l'allèle à risque parmi les sujets séquencés : s'il est absent dans les données de séquence, il ne pourra pas être imputé et restera donc absent dans l'ensemble des données GWAS. La recherche de variants rares nécessite donc des effectifs importants. Cependant, en sélectionnant des sujets malades pour l'étape de séquençage, on peut augmenter la proportion des porteurs d'allèles à risque et donc la probabilité de retrouver ces allèles dans les données GWAS après imputations.

Conclusion et discussion

Cette thèse a eu pour objectifs d'évaluer différentes méthodes statistiques et designs d'étude pour la recherche d'association à grande échelle. Ces travaux ont été développés dans le cadre de l'analyse de données pan-génomiques de la maladie de Parkinson, dans le but d'améliorer notre compréhension de la composante génétique de cette pathologie en particulier, mais aussi d'autres maladies multifactorielles, comme la sclérodermie.

Durant ma thèse, nous nous sommes intéressés principalement aux études d'association pan-génomiques dans deux contextes : la recherche d'association de variations communes et de variations rares par l'approche simple-marqueur et par l'approche multi-marqueur sur des données à grande échelle de génotypes, de pseudo-séquences et de séquences.

Nous avons conçu une étude d'association pan-génomique de la MP, en trois étapes sur des données Françaises et Australiennes. Notre étude a été fructueuse et nous avons réussi à confirmer l'association avec deux gènes connus de la MP et identifier deux autres loci. Le premier gène avait déjà été identifié dans une étude GWA Japonaise, mais jamais dans la population Européenne. Le deuxième est identifié pour la première fois par notre étude. Ce gène se situe à 20kb d'un autre gène contenant des polymorphismes impliqués dans la maladie bipolaire. Cependant, son rôle dans la MP reste inconnu.

Pour augmenter la puissance des études GWAS individuelles, et détecter d'autres associations, nous avons réalisé une étude de méta-analyse dans le cadre de l'IPDGC "International Parkinson's Disease Genomic Consortium", qui inclue des données Nord-américaines, Allemandes, Hollandaises, Anglaises, Islandaises et nos données Françaises. Nous avons procédé en deux étapes. Dans la première étape, nous avons imputé les génotypes manquants dans les puces de GWASs pour augmenter la couverture génétique. Dans la seconde, nous avons combiné les résultats des GWASs individuelles par des méthodes statistiques de méta-analyse.

Cette étude a permis d'établir, de façon définitive, l'implication de six gènes déjà connus de la MP (SNCA, MAPT, BST1, LRRK2, GAK et HLA-DRB5) et d'identifier plusieurs variants dans cinq nouveaux gènes/loci (ACMSD, STK39, MCCC1/LAMP3, SYT11, CCDC62/HIP1R).

Cette étude a été suivie d'une nouvelle méta-analyse des données de l'IPDGC adossée sur des données de 23andMe : elle a réussi à identifier des variants associés à la MP de sept nouveaux gènes (RAB7L1/PARK16, NMD3STBD1, GPNMB, FGF20, MMP16, STX1B).

Au total, près d'une vingtaine de locus contribuant au risque de la MP ont pu être identifiés par ces deux méta-analyses. La grande majorité des locus sont nouveaux. Il y a trois ans, leur implication dans le risque de la MP n'était pas connue. Les études pan-génomiques confirment aussi l'existence d'un continuum entre les formes monogéniques et celles communes de la MP. Cependant, les variants, à ce jour, identifiés ont des effets faibles sur le risque de la maladie et n'expliquent qu'une faible part des cas. Il est clair que la variabilité ponctuelle n'est pas le seul type de variabilité du génome. Les variations structurelles comme les variations du nombre de copies (CNV) ou de l'épi-génétique échappent aux études d'association pan-génomiques. Par ailleurs, des facteurs de l'environnement, seuls ou en interactions avec des facteurs génétiques, peuvent aussi expliquer une partie non négligeable du risque de la maladie.

Pendant ma thèse, nous nous sommes limités à la caractérisation du risque expliqué par la variabilité ponctuelle de l'ADN. Dans ce contexte, il est important de noter que les résultats GWASs sont plus probablement obtenus sous l'association indirecte que directe : les études GWAS identifient des signaux d'associations et non le(s) variant(s) causal(ux).

Idéalement, pour que l'association soit directe il faut observer le variant causal. Ceci requiert d'avoir la séquence entière du génome de plusieurs milliers de malades et témoins, ce qui reste impossible à cause du coût prohibitif de l'étude. Pour augmenter la probabilité de l'association directe on peut, alternativement, améliorer la couverture de la variabilité génétique locale en utilisant conjointement plusieurs marqueurs génétiques par des tests de multi-marqueur.

Nous avons comparé deux tests d'association multi-marqueur: le test haplotypique et le test « SNP-set ». Le test haplotypique présente plus de défis, tant au niveau calcul, à cause de l'inférence des phases alléliques, qu'au niveau du nombre de *df*, qui augmente avec le nombre de SNPs inclus dans le modèle. Nous avons cherché à quantifier le gain apporté par ces approches vis-à-vis de l'approche simple-marqueur, corrigée par le nombre de SNPs. Les conclusions principales de notre étude sont les suivantes: Nous avons observé peu de différences dans les distributions de ces deux tests. En conséquence, du fait de sa simplicité de calcul, le test « SNP-set » peut-être proposé comme une première approche rapide d'analyse

jointe de plusieurs marqueurs ; L'avantage théorique de la réduction du nombre de degrés de liberté du test « SNP-set » n'apparait pas dans nos données de la MP ; Globalement et dans nos données, les tests multi-marqueur n'apparaissent pas très avantageux, par rapport au test simple-marqueur. Notons, cependant que pour un certain nombre de gènes, l'évidence statistique de l'association est bien meilleure sous le test haplotypique, prenant en compte les blocs de LD du gène, que sous le test simple-marqueur.

Une autre hypothèse peut expliquer la part restante du risque de la MP, celle de l'implication de variants peu fréquents, voir rares, à effets importants. Différents tests d'association ont été proposés comme alternatifs au test simple-marqueur non-puissant pour l'analyse de variants rares. Leur approche générale est de combiner les allèles rares des variants d'une même unité génomique (le gène) en une seule variable. Les extensions ont été introduites pour filtrer ou non les allèles à combiner et/ou pour pondérer ou non les contributions individuelles des allèles rares du gène. Nous avons évalué la performance (erreurs de type I et de type II) de plusieurs méthodes statistiques dans des données de séquences. Cette étude a été faite dans le cadre d'un atelier de travail intitulé « Genetic Analysis Workshop 17». Parmi les conclusions principales que nous avons tirées:

1) Nous n'avons pas identifié d'approche qui soit globalement meilleure que les autres. Les performances relatives des méthodes variaient selon le gène testé.

2) Pour la détection d'association de variants rares, la puissance de l'analyse d'association en données familiales peut s'avérer nettement plus élevée que celle obtenue en données de population. La performance du design familial est d'autant plus importante que les familles sont grandes. De plus, notre évaluation a été faite dans un échantillon de familles où la distribution du phénotype était aléatoire. On peut s'attendre à ce que la performance de l'analyse d'association dans des données familiales soit encore meilleure si les familles sont sélectionnées sur des sujets ayant des valeurs extrêmes au traité étudié. L'évaluation de ces méthodes dans des données sélectionnées serait intéressante à développer.

3) Finalement, les approches de type « collapsing » ont une meilleure puissance par rapport au test simple-marqueur. En revanche, les niveaux de puissance que nous avons observés sont relativement petits, même au seuil de signification de 5%. Pour des critères de signification stricts (i.e. 1.6×10^{-6} = 0.05/30000 gènes), les niveaux de puissance seront évidement plus faibles.

Pour avoir une bonne puissance (>80%), il faut donc de grandes tailles d'échantillons (plusieurs dizaines de milliers de sujets). Ainsi, le coût du séquençage du génome ou de l'exome entier d'un grand nombre de sujets reste l'entrave essentielle à la recherche d'association de variants rares. Un design d'étude alternatif, appelé pseudo-séquençage, a été récemment proposé. Il repose sur la combinaison de données publiques de séquences aux données génotypiques de GWAS à travers des techniques d'imputation. Un de nos projets est de conduire une méta-analyse, basée sur ce design, dans les données de l'IPDGC pour la recherche de variants rares de MP. Le test d'association utilisé est celui du modèle de la fonction de noyau, limité aux variants peu fréquents (MAF<5%).

Références

1. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., . . . Zhu, X. The sequence of the human genome. *Science* 291, 1304-51 (2001).
2. Zaykin, D.V., Westfall, P.H., Young, S.S., Karnoub, M.A., Wagner, M.J. & Ehm, M.G. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum Hered* 53, 79-91 (2002).
3. Excoffier, L. & Slatkin, M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 12, 921-7 (1995).
4. Stephens, M., Smith, N.J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 68, 978-89 (2001).
5. Lewontin, R.C.K., K. The evolutionary dynamics of complex polymorphisms. *Evolution* 14, 458–472 (1960).
6. Hedrick, P.W. Gametic disequilibrium measures: proceed with caution. *Genetics* 117, 331-41 (1987).
7. Lewontin, R.C. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics* 49, 49-67 (1964).
8. Dawson, E., Abecasis, G.R., Bumpstead, S., Chen, Y., Hunt, S., Beare, D.M., . . . Dunham, I. A first-generation linkage disequilibrium map of human chromosome 22. *Nature* 418, 544-8 (2002).
9. Price, A.L., Weale, M.E., Patterson, N., Myers, S.R., Need, A.C., Shianna, K.V., . . . Reich, D. Long-range LD can confound genome scans in admixed populations. *Am J Hum Genet* 83, 132-5; author reply 135-9 (2008).
10. Haseman, J.K. & Elston, R.C. The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* 2, 3-19 (1972).
11. Cochran, W. Some methods for strengthening the common chi-squared tests. *Biometrics (International Biometric Society)* 10, 417-451 (1954).
12. Armitage, P. Tests for Linear Trends in Proportions and Frequencies. *Biometrics (International Biometric Society)* 11, 375–386 (1955).
13. Agresti, A. *Categorical Data Analysis*, (2002).
14. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., . . . Sham, P.C. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81, 559-75 (2007).
15. Balding, D.J. A tutorial on statistical methods for population association studies. *Nat Rev Genet* 7, 781-91 (2006).
16. Cardon, L.R. & Palmer, L.J. Population stratification and spurious allelic association. *Lancet* 361, 598-604 (2003).
17. Wang, W.Y., Barratt, B.J., Clayton, D.G. & Todd, J.A. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 6, 109-18 (2005).
18. Maraganore, D.M., de Andrade, M., Lesnick, T.G., Strain, K.J., Farrer, M.J., Rocca, W.A., . . . Ballinger, D.G. High-resolution whole-genome association study of Parkinson disease. *Am J Hum Genet* 77, 685-93 (2005).
19. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* 273, 1516-7 (1996).
20. D. Altsuler, L.D.B., A. Chakravarti, et al. A haplotype map of the human genome. *Nature* 437, 1299-320 (2005).
21. Li, M., Li, C. & Guan, W. Evaluation of coverage variation of SNP chips for genome-wide association studies. *Eur J Hum Genet* 16, 635-43 (2008).

22. Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., . . . Hoh, J. Complement factor H polymorphism in age-related macular degeneration. *Science* 308, 385-9 (2005).
23. Pakkenberg, B., Moller, A., Gundersen, H.J., Mouritzen Dam, A. & Pakkenberg, H. The absolute number of nerve cells in substantia nigra in normal subjects and in patients with Parkinson's disease estimated with an unbiased stereological method. *J Neurol Neurosurg Psychiatry* 54, 30-3 (1991).
24. Calne, D. A definition of Parkinson's disease. *Parkinsonism Relat Disord* 11 Suppl 1, S39-40 (2005).
25. Gelb, D.J., Oliver, E. & Gilman, S. Diagnostic criteria for Parkinson disease. *Arch Neurol* 56, 33-9 (1999).
26. Hughes, A.J., Daniel, S.E., Kilford, L. & Lees, A.J. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. *J Neurol Neurosurg Psychiatry* 55, 181-4 (1992).
27. von Campenhausen, S., Bornschein, B., Wick, R., Botzel, K., Sampaio, C., Poewe, W., . . . Dodel, R. Prevalence and incidence of Parkinson's disease in Europe. *Eur Neuropsychopharmacol* 15, 473-90 (2005).
28. Samii, A., Nutt, J.G. & Ransom, B.R. Parkinson's disease. *Lancet* 363, 1783-93 (2004).
29. Twelves, D., Perkins, K.S. & Counsell, C. Systematic review of incidence studies of Parkinson's disease. *Mov Disord* 18, 19-31 (2003).
30. de Lau, L.M. & Breteler, M.M. Epidemiology of Parkinson's disease. *Lancet Neurol* 5, 525-35 (2006).
31. Elbaz, A., Clavel, J., Rathouz, P.J., Moisan, F., Galanaud, J.P., Delemotte, B., . . . Tzourio, C. Professional exposure to pesticides and Parkinson disease. *Ann Neurol* 66, 494-504 (2009).
32. Payami, H., Zareparsi, S., James, D. & Nutt, J. Familial aggregation of Parkinson disease: a comparative study of early-onset and late-onset disease. *Arch Neurol* 59, 848-50 (2002).
33. Sveinbjornsdottir, S., Hicks, A.A., Jonsson, T., Petursson, H., Gugmundsson, G., Frigge, M.L., . . . Stefansson, K. Familial aggregation of Parkinson's disease in Iceland. *N Engl J Med* 343, 1765-70 (2000).
34. Lesage, S. & Brice, A. Parkinson's disease: from monogenic forms to genetic susceptibility factors. *Hum Mol Genet* 18, R48-59 (2009).
35. Poorkaj, P., Bird, T.D., Wijsman, E., Nemens, E., Garruto, R.M., Anderson, L., . . . Schellenberg, G.D. Tau is a candidate gene for chromosome 17 frontotemporal dementia. *Ann Neurol* 43, 815-25 (1998).
36. Sidransky, E. Gaucher disease: complexity in a "simple" disorder. *Mol Genet Metab* 83, 6-15 (2004).
37. Fung, H.C., Scholz, S., Matarin, M., Simon-Sanchez, J., Hernandez, D., Britton, A., . . . Singleton, A. Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol* 5, 911-6 (2006).
38. Myers, R.H. Considerations for genomewide association studies in Parkinson disease. *Am J Hum Genet* 78, 1081-2 (2006).
39. Satake, W., Nakabayashi, Y., Mizuta, I., Hirota, Y., Ito, C., Kubo, M., . . . Toda, T. Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. *Nat Genet* 41, 1303-7 (2009).

40. Simon-Sanchez, J., Schulte, C., Bras, J.M., Sharma, M., Gibbs, J.R., Berg, D., . . . Gasser, T. Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat Genet* 41, 1308-12 (2009).
41. Spencer, C.C., Plagnol, V., Strange, A., Gardner, M., Paisan-Ruiz, C., Band, G., . . . Wood, N.W. Dissection of the genetics of Parkinson's disease identifies an additional association 5' of SNCA and multiple associated haplotypes at 17q21. *Hum Mol Genet* 20, 345-53 (2011).
42. Vascular factors and risk of dementia: design of the Three-City Study and baseline characteristics of the study population. *Neuroepidemiology* 22, 316-25 (2003).
43. WTCCC2. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661-78 (2007).
44. Moskvina, V., Craddock, N., Holmans, P., Owen, M.J. & O'Donovan, M.C. Effects of differential genotyping error rate on the type I error probability of case-control studies. *Hum Hered* 61, 55-64 (2006).
45. Kong, A., Masson, G., Frigge, M.L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., . . . Stefansson, K. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet* 40, 1068-75 (2008).
46. Miyazawa, H., Kato, M., Awata, T., Kohda, M., Iwasa, H., Koyama, N., . . . Hagiwara, K. Homozygosity haplotype allows a genomewide search for the autosomal segments shared among patients. *Am J Hum Genet* 80, 1090-102 (2007).
47. Houwen, R.H., Baharloo, S., Blankenship, K., Raeymaekers, P., Juyn, J., Sandkuijl, L.A. & Freimer, N.B. Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nat Genet* 8, 380-6 (1994).
48. Te Meerman, G.J., Van der Meulen, M.A. & Sandkuijl, L.A. Perspectives of identity by descent (IBD) mapping in founder populations. *Clin Exp Allergy* 25 Suppl 2, 97-102 (1995).
49. Morris, E.Z.a.A. *Analysis of Complex Disease Association Studies A pratical guide*, (2011).
50. Marchini, J., Cardon, L.R., Phillips, M.S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nat Genet* 36, 512-7 (2004).
51. Freedman, M.L., Reich, D., Penney, K.L., McDonald, G.J., Mignault, A.A., Patterson, N., . . . Altshuler, D. Assessing the impact of population stratification on genetic association studies. *Nat Genet* 36, 388-93 (2004).
52. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* 55, 997-1004 (1999).
53. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* 155, 945-59 (2000).
54. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. & Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38, 904-9 (2006).
55. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* 2, e190 (2006).
56. Menozzi, P., Piazza, A. & Cavalli-Sforza, L. Synthetic maps of human gene frequencies in Europeans. *Science* 201, 786-92 (1978).
57. Saad, M., Lesage, S., Saint-Pierre, A., Corvol, J.C., Zelenika, D., Lambert, J.C., . . . Brice, A. Genome-wide association study confirms BST1 and suggests a locus on 12q24 as the risk loci for Parkinson's disease in the European population. *Hum Mol Genet* 20, 615-27 (2011).

58. Healy, D.G., Abou-Sleiman, P.M., Lees, A.J., Casas, J.P., Quinn, N., Bhatia, K., . . . Wood, N.W. Tau gene and Parkinson's disease: a case-control study and meta-analysis. *J Neurol Neurosurg Psychiatry* 75, 962-5 (2004).
59. Zhang, J., Song, Y., Chen, H. & Fan, D. The tau gene haplotype h1 confers a susceptibility to Parkinson's disease. *Eur Neurol* 53, 15-21 (2005).
60. Wilson, C.J. & Callaway, J.C. Coupled oscillator model of the dopaminergic neuron of the substantia nigra. *J Neurophysiol* 83, 3084-100 (2000).
61. Simon-Sanchez, J., van Hilten, J.J., van de Warrenburg, B., Post, B., Berendse, H.W., Arepalli, S., . . . Heutink, P. Genome-wide association study confirms extant PD risk loci among the Dutch. *Eur J Hum Genet* 19, 655-61 (2011).
62. Nalls, M.A., Plagnol, V., Hernandez, D.G., Sharma, M., Sheerin, U.M., Saad, M., . . . Wood, N.W. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet* 377, 641-9 (2011).
63. Glaser, B., Kirov, G., Bray, N.J., Green, E., O'Donovan, M.C., Craddock, N. & Owen, M.J. Identification of a potential bipolar risk haplotype in the gene encoding the winged-helix transcription factor RFX4. *Mol Psychiatry* 10, 920-7 (2005).
64. Hamza, T.H., Zabetian, C.P., Tenesa, A., Laederach, A., Montimurro, J., Yearout, D., . . . Payami, H. Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. *Nat Genet* 42, 781-5 (2010).
65. Zaykin, D.V. Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *J Evol Biol* 24, 1836-41 (2011).
66. Stouffer, S., DeVinney, L. & Suchmen, E. The American Soldier: Adjustement During Army Life. *Princeton University Press, Princeton, NJ* 1(1949).
67. Lipták, T. On the combination of independent tests. *Magyar Tud. Akad. Mat. Kutato Int. Közel* 3, 171-196 (1958).
68. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26, 2190-1 (2010).
69. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu Rev Genomics Hum Genet* 10, 387-406 (2009).
70. Li, Y., Ding, J. & Abecasis, G. Mach 1.0: rapid haplotype reconstruction and missing genotype inference. *Am. J. Hum. Genet* 79:S2290(2006).
71. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 39, 906-13 (2007).
72. Browning, S.R. Multilocus association mapping using variable-length Markov chains. *Am J Hum Genet* 78, 903-13 (2006).
73. IPDGC. A two-stage meta-analysis identifies several new loci for Parkinson's disease. *PLoS Genet* 7, e1002142 (2011).
74. Do, C.B., Tung, J.Y., Dorfman, E., Kiefer, A.K., Drabant, E.M., Francke, U., . . . Eriksson, N. Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. *PLoS Genet* 7, e1002141 (2011).
75. Schaid, D.J., McDonnell, S.K., Hebbring, S.J., Cunningham, J.M. & Thibodeau, S.N. Nonparametric tests of association of multiple genes with human disease. *Am J Hum Genet* 76, 780-93 (2005).
76. Kwee, L.C., Liu, D., Lin, X., Ghosh, D. & Epstein, M.P. A powerful and flexible multilocus association test for quantitative traits. *Am J Hum Genet* 82, 386-97 (2008).

77. Wu, M.C., Kraft, P., Epstein, M.P., Taylor, D.M., Chanock, S.J., Hunter, D.J. & Lin, X. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet* 86, 929-42 (2010).
78. Wang, T. & Elston, R.C. Improved power by use of a weighted score test for linkage disequilibrium mapping. *Am J Hum Genet* 80, 353-60 (2007).
79. Pan, W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet Epidemiol* 33, 497-507 (2009).
80. Johnson, N. & Kotz, S. Distributions in Statistics, Continuous Univariate Distributions. *Boston: Houghton-Mifflin* (1970).
81. Zhang, J.-T. Approximate and Asymptotic Distributions of Chi-Squared-Type Mixtures With Applications. *Journal of the American Statistical Association* 100(2005).
82. Goeman JJ, Van de Geer S & Van Houwelingen HC. Testing against a high dimensional alternative. *J Royal Stat Soc B* 68, 477–493 (2006).
83. Liu, D., Ghosh, D. & Lin, X. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics* 9, 292 (2008).
84. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M. & Lin, X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 89, 82-93 (2011).
85. Neale, B.M., Rivas, M.A., Voight, B.F., Altshuler, D., Devlin, B., Orho-Melander, M., . . . Daly, M.J. Testing for an unusual distribution of rare variants. *PLoS Genet* 7, e1001322 (2011).
86. Neyman J & E, S. On the use of c(α) optimal tests of composite hypotheses. *Bulletin of the International Statistical Institute*, 477–497 (1966).
87. Davies, R. The distribution of a linear combination of chi-square random variables. *J. R. Stat. Soc. Ser. C Appl. Stat.* 29, 323–333 (1980).
88. Li, Y., Byrnes, A.E. & Li, M. To identify associations with rare variants, just WHaIT: Weighted haplotype and imputation-based tests. *Am J Hum Genet* 87, 728-35 (2010).
89. Iyengar, S.K. & Elston, R.C. The genetic basis of complex traits: rare variants or "common gene, common disease"? *Methods Mol Biol* 376, 71-84 (2007).
90. Almasy, L., Dyer, T.D., Peralta, J.M., Kent, J.W., Jr., Charlesworth, J.C., Curran, J.E. & Blangero, J. Genetic Analysis Workshop 17 mini-exome simulation. *BMC Proc* 5 Suppl 9, S2 (2011).
91. Ziegler, A., Ghosh, S., Dyer, T.D., Blangero, J., MacCluer, J. & Almasy, L. Introduction to genetic analysis workshop 17 summaries. *Genet Epidemiol* 35 Suppl 1, S1-4 (2011).
92. Morgenthaler, S. & Thilly, W.G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res* 615, 28-56 (2007).
93. Li, B. & Leal, S.M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 83, 311-21 (2008).
94. Frazer, K.A., Murray, S.S., Schork, N.J. & Topol, E.J. Human genetic variation and its contribution to complex traits. *Nat Rev Genet* 10, 241-51 (2009).
95. Gorlov, I.P., Gorlova, O.Y., Sunyaev, S.R., Spitz, M.R. & Amos, C.I. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am J Hum Genet* 82, 100-12 (2008).
96. Morris, A.P. & Zeggini, E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* 34, 188-93 (2010).

97. Bansal, V., Libiger, O., Torkamani, A. & Schork, N.J. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* 11, 773-85 (2010).
98. Madsen, B.E. & Browning, S.R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 5, e1000384 (2009).
99. Price, A.L., Kryukov, G.V., de Bakker, P.I., Purcell, S.M., Staples, J., Wei, L.J. & Sunyaev, S.R. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 86, 832-8 (2010).
100. Zawistowski, M., Gopalakrishnan, S., Ding, J., Li, Y., Grimm, S. & Zollner, S. Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. *Am J Hum Genet* 87, 604-17 (2010).

Human Molecular Genetics, 2011, Vol. 20, No. 3 **615–627**
doi:10.1093/hmg/ddq497
Advance Access published on November 17, 2010

Genome-wide association study confirms BST1 and suggests a locus on 12q24 as the risk loci for Parkinson's disease in the European population

Mohamad Saad[1,2,†], Suzanne Lesage[3,4,5,†], Aude Saint-Pierre[1,2], Jean-Christophe Corvol[3,4,5,6], Diana Zelenika[7], Jean-Charles Lambert[8,9,10], Marie Vidailhet[3,4,5], George D. Mellick[11,12], Ebba Lohmann[3,4,5], Franck Durif[13], Pierre Pollak[14], Philippe Damier[15], François Tison[16], Peter A. Silburn[11,12], Christophe Tzourio[17,18], Sylvie Forlani[3,4,5], Marie-Anne Loriot[19,20,21], Maurice Giroud[22], Catherine Helmer[23], Florence Portet[24], Philippe Amouyel[8,9,10,25], Mark Lathrop[7], Alexis Elbaz[17,18], Alexandra Durr[3,4,5,26], Maria Martinez[1,2,*] and Alexis Brice[3,4,5,26,*] for the French Parkinson's Disease Genetics Study Group[‡]

[1]INSERM U563, CPTP, CHU Purpan, 31024 Toulouse, France, [2]Paul Sabatier University, Toulouse, France, [3]Université Pierre et Marie Curie-Paris6, Centre de Recherche de l'Institut du Cerveau et de la Moelle épinière, UMR-S975, Paris, France, [4]INSERM U975, Paris, France, [5]CNRS, UMR 7225, Paris, France, [6]INSERM CIC-9503, Hôpital Pitié-Salpêtrière, Paris, France, [7]Centre National de Génotypage, Institut Génomique, Commissariat à l'Energie Atomique, Evry, France, [8]INSERM U744, Lille, France, [9]Institut Pasteur de Lille, Lille, France, [10]Université de Lille Nord, Lille, France, [11]National Centre for Adult Stem Cell Research, Eskitis Institute for Cell and Molecular Therapies, Griffith University, Brisbane, Queensland, Australia, [12]Department of Neurology, Princess Alexandra Hospital, Brisbane, Queensland, Australia, [13]Service de Neurologie, Hôpital Gabriel Montpied, Clermont-Ferrand, France, [14]Service de Neurologie, CHU de Grenoble, Grenoble, France, [15]CHU Nantes, CIC0004, Service de Neurologie, Nantes, France, [16]Service de Neurologie, Hôpital Haut-Lévêque, Pessac, France, [17]INSERM U708, Paris, France, [18]Université Pierre et Marie Curie Paris6, Paris, France, [19]UMR-S775, Paris, France, [20]Université Paris Descartes, Paris, France, [21]AP-HP, Hôpital Européen Georges Pompidou, Paris, France, [22]Centre Hospitalier Dijon, Dijon, France, [23]INSERM, CR897, Université Victor Segalen Bordeaux-2, Bordeaux, France, [24]INSERM U888, Montpellier, France, [25]CHRU de Lille, Lille, France, and [26]Department of Genetics and Cytogenetics, AP-HP, Pitié-Salpêtrière Hospital, Paris, France

Received July 16, 2010; Revised and Accepted November 10, 2010

We performed a three-stage genome-wide association study (GWAS) to identify common Parkinson's disease (PD) risk variants in the European population. The initial genome-wide scan was conducted in a French sample of 1039 cases and 1984 controls, using almost 500 000 single nucleotide polymorphisms (SNPs). Two SNPs at SNCA were found to be associated with PD at the genome-wide significance level ($P < 3 \times 10^{-8}$). An additional set of promising and new association signals was identified and submitted for immediate replication in two independent case–control studies of subjects of European descent. We first carried out an *in silico* replication study using GWAS data from the WTCCC2 PD study sample (1705 cases, 5200 WTCCC controls). Nominally replicated SNPs were further genotyped in a third sample of 1527 cases and 1864 controls from France and Australia. We found converging evidence of association with PD on 12q24 (rs4964469,

*To whom correspondence should be addressed. Tel: +33 562744587; Fax: +33 562744558; Email: maria.martinez@inserm.fr (M.M.); Tel: +33 142162189; Fax: +33 144243658; Email: alexis.brice@upmc.fr (A.B.)
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
‡The French Parkinson's Disease Genetics Study Group includes Y. Agid, M. Anheim, A.-M. Bonnet, M. Borg, A. Brice, E. Broussolle, J.-C. Corvol, Ph. Damier, A. Destée, A. Dürr, F. Durif, S. Klebe, E. Lohmann, M. Martinez, C. Penet, P. Pollak, O. Rascol, F. Tison, C. Tranchant, M. Vérin, F. Viallet and M. Vidailhet.

combined $P = 2.4 \times 10^{-7}$) and confirmed the association on 4p15/BST1 (rs4698412, combined $P = 1.8 \times 10^{-6}$), previously reported in Japanese data. The 12q24 locus includes RFX4, an isoform of which, named *RFX4_v3*, encodes brain-specific transcription factors that regulate many genes involved in brain morphogenesis and intracellular calcium homeostasis.

INTRODUCTION

Parkinson's disease (PD) is the second most common degenerative disease, affecting 1–2% of individuals older than 65 years. Clinical features of PD result primarily from the loss of dopaminergic neurons in the substantia nigra. Although the common form of PD is sporadic, six genes have been identified, mainly by linkage analyses of Mendelian forms of the disease. Two genes, SNCA (encoding α-synuclein) and LRRK2, have an autosomal dominant inheritance and four other genes, PARK2 (parkin), PARK6 (PINK1), PARK7 (DJ-I) and PARK13 (ATP13A2), have an autosomal recessive inheritance (1). Frequently, mutations in these genes are found in patients with early-onset PD, particularly those with autosomal recessive inheritance. However, in most populations, Mendelian forms of parkinsonism are rare when compared with the most common form of PD, a frequent and complex disorder probably explained by the interaction between genetic and environmental factors.

The first two genome-wide association studies (GWASs) in PD (2,3) provided evidence of association with several loci but most often not at the genome-wide significant level, and most initial association findings were not confirmed by subsequent replication analyses (4). Two recent GWASs (5,6) reported strong or genome-wide significant associations with one or more of the known PD genes (i.e. SNCA, MAPT and/or LRRK2). So far, only two 'new' loci have been identified, 1q32/PARK16 and 4p15/BST1, in the Japanese data (5). The US/UK/German GWAS (6) replicated positive association with variants at PARK16 but failed to replicate the association at BST1.

To identify additional variants that affect PD risk in the European population, we designed a three-stage GWAS of PD in three case–control samples from France, the UK and Australia (total of >13 300 subjects). A set of 50 top association signals was identified in the scan sample (1039 cases and 1984 controls from France) using the Illumina-610Quad chip. Promising and new signals were followed-up for stepwise replication in two further UK and French/Australian case–control studies (>3200 cases and 7000 controls).

RESULTS

The genome-wide association results from logistic test corrected for genomic inflation (GC) revealed two single nucleotide polymorphisms (SNPs) with $P_{GC} < 10^{-7}$, and a substantial number of SNPs with strong ($P_{GC} < 10^{-4}$) evidence of association (Fig. 1 and Table 1). For practical reasons, we focus our attention on the 50 best-associated SNPs to prioritize for immediate *in silico* replication (Table 1). Secondary logistic analyses, adjusted for the first two principal componenets (PCs) led to similar rank order of SNPs, albeit slightly weaker association signals (Table 1). This suggests that the significant results, revealed by our primary analyses, are not biased by residual population substructure within our French scan sample. The 50 best-associated SNPs spanned 23 distinct genomic loci, and were associated with $P_{GC} < 5.6 \times 10^{-5}$. Sixteen associations were found within two well-known PD genes, i.e. SNCA (4q22, 4 SNPs) and MAPT (17q12–q21, 11 SNPs), or within BST1 (4p15, one SNP), a recently reported PD risk locus established at the genome-wide significance level in a Japanese population (5). The remaining 34 SNPs were located in 20 distinct previously unreported putative PD loci. The two genome-wide significant SNPs were located on 4q22/SNCA [rs356220, $P_{GC} = 2.82 \times 10^{-8}$, OR = 1.37; 95% CI (1.22–1.53) and rs2736990, $P_{GC} = 2.88 \times 10^{-8}$, OR = 1.35 (1.22–1.50)]. The next most significant SNP was on chromosome 12q21/LOC401725 [rs7954761, $P_{GC} = 2.09 \times 10^{-7}$, OR = 1.34 (1.20–1.50)].

The 50 top SNPs were tested for *in silico* replication in the WTCCC2 PD study data (Table 1). For the sake of clarity, OR values are reported as a function of the number of risk alleles as identified in the stage-1 data. Associations for all 15 SNPs in SNCA and MAPT genes were replicated at nominal P-values $< 4 \times 10^{-5}$. Association with the BST1 variant was also replicated but at a weaker significance level (OR = 1.08, $P = 0.025$). For all SNPs at SNCA, MAPT and BST1, the results in the French scan and the UK replication samples were highly congruent in terms of risk alleles and allele frequencies. As expected, ORs estimated in our scan study tend to be higher than those obtained in the replication-stage data, especially for BST1. Of the remaining 20 loci, association signals at three loci (four SNPs) were replicated with nominal $P < 5\%$ and with the same direction of effect. These SNPs were located on chromosomes 2q21.3 (rs621341, OR = 1.08, $P = 0.028$ and rs6723108, OR = 1.11, $P = 0.005$), 12p13.3 (rs11064524, OR = 1.08, $P = 0.045$) and 12q24 (rs4964469, OR = 1.11, $P = 0.0045$). Differences in allele frequencies across the data from France and UK were notable for the 2q21–q22 SNPs. Indeed, the region encompasses the LCT (lactase) gene whose SNPs are known to vary in frequency across Europe, and rs6723108 has been shown to have different allele frequencies in the French and the UK–Irish populations (7).

We further followed-up the five replicated SNPs (from three newly identified loci and from BST1) in the second replication dataset (1527 cases and 1864 controls from France and Australia) (Table 2). In stage 3, evidence of association was assessed with the Mantel–Haenszel test to control for the potential confounding owing to the different geographical origins (France versus Australia). Evidence of association was replicated for two SNPs located on 12q24 (rs4964469, OR = 1.12, $P = 0.0175$) and on 4p15/BST1

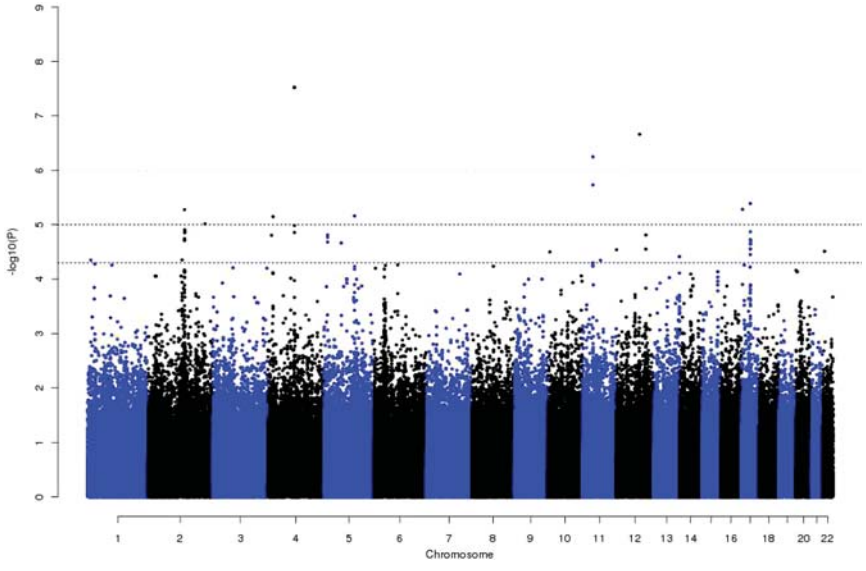


Figure 1. Manhattan plot of the genome-wide association results for 492 929 SNPs. Logistic analysis corrected for genomic inflation (GC results).

(rs4698412, OR = 1.10, $P = 0.029$). Association signals from joint analysis of the two replication datasets were improved for the same two SNPs only: at 4p15/BST1 (stage 2 + stage 3, $P = 0.0033$) and at 12q24 (stage 2 + stage 3, $P = 0.00036$) loci. Notably, joint analysis of the three datasets showed a consistently greater support for association for the newly identified locus on chromosome 12q24 ($P = 2.38 \times 10^{-7}$) than for 4p15/BST1 ($P = 1.79 \times 10^{-6}$). Additional analyses showed that our initial association signals were not confounded by age and they did not appear to be driven either by the subgroup of cases having an early age (<50) of onset of the disease or by those having a positive family history of PD (results not shown). The population-attributable risk (PAR) associated with SNCA, MAPT, BST1 and the 12q24 locus estimated in stage-1 data was 11, 20, 13 and 8%, respectively; in the combined data, PAR was 7% and 4% for BST1 and 12q24, respectively.

Finally, we also examined 18 SNPs from five loci, previously reported to be associated with PD at a genome-wide significance level from two published GWASs of PD (5,6) (Table 3). We added three SNPs (at SNCA and BST1) that were found strongly associated in our stage-1 data. The table also shows the results for a suggestive PD risk locus (GAK) reported by the published GWAS of PD from familial cases (8). As for the previously reported PD loci, two loci only (SNCA and MAPT) have been identified with genome-wide significance at the screen stage: SNCA in both the Japanese and European populations and MAPT in the European population only. The two new PD risk loci (BST1 and PARK16) were identified in the Japanese population: association signals were strong ($P < 10^{-6}$) in the discovery sample, and exceeded $P < 10^{-8}$ in the combined data (5). As already reported here, our GWAS provided genome-wide significance for two SNCA variants and replicated positive associations for variants at MAPT and BST1 loci. It is worth noting that allele frequencies may differ markedly between the Japanese and the European datasets, especially for SNPs at the SNCA and BST1 loci. Saliently, the directions of effects (i.e. risk allele) and effect sizes at SNCA variants are rather congruent across the European and Japanese datasets. For the remaining three PD loci, evidence of association was nominal (PARK16, $P_{GC} = 0.03$; LRKK2, $P_{GC} = 0.04$; GAK $P_{GC} = 0.008$) in the France-GWAS data.

DISCUSSION

Our genome-wide association analyses in the French scan data revealed two SNPs with genome-wide significance ($P_{GC} < 10^{-7}$), and a number of additional SNPs with suggestive evidence ($P_{GC} < 10^{-4}$). Here, we focused on the 50 top associated SNPs for immediate replication in two independent case–control samples. We used a stepwise replication design. To refine the set of most promising results, we first conducted *in silico* replication for the 50 SNPs in the WTCCC2 PD data (1705 cases and 5200 WTCCC controls). Replicated SNPs were genotyped and tested in a third dataset of 1527 cases and 1864 controls from France and Australia. Our scan stage showed genome-wide significance of association with PD for two SNPs at the 4q22/SNCA locus ($P_{GC} < 2.88 \times 10^{-8}$). Indeed, out of the 50 top associated SNPs, 15 are located in genomic regions of two known PD genes (SCNCA, MAPT) and one is located on 4p15/BST1, a risk locus recently reported with genome-wide significance in Japanese samples. SNPs at SNCA and MAPT were all significantly associated with PD in the UK-GWAS data (SNCA, $P < 8 \times 10^{-5}$; MAPT, $P < 2.75 \times 10^{-6}$). Evidence of association with 4p15/BST1 was also replicated in the UK sample but at a lower (rs4698412, $P = 0.025$) significance level. Out of the remaining 34 SNPs, four SNPs (three loci) showed significant ($P < 0.05$) and consistent evidence of association in the UK data. A total of five SNPs (four loci: 2q21.3, 12p13.3, 12q24 and BST1) were followed-up for replication in the third case–control sample. Two of the four tested regions were replicated: 4p15/BST1 ($P = 0.03$) and 12q24

Table 1. Top 50 SNPs in scan stage and *in silico* replication results

Chromosome (gene)	Position (bp)	SNP	Stage-1: scan (France) data					Stage-2: replication (UK) data		
			RA[a]	RAF[b]	OR	P_{GC} (two-tailed)[c]	P_{2PCs} (two-tailed)[d]	RAF	OR	P (one-tailed)[e]
Known PD genes/previously published loci										
4q22 (SNCA)	90858538	rs11931074	T	0.07	1.52	1.35E-05	9.04E-05	0.07	1.33	4.01E-05
	90860363	rs356220	T	0.35	1.37	2.82E-08	6.26E-07	0.36	1.27	2.59E-09
	90894261	rs3857059	G	0.07	1.54	1.00E-05	6.32E-05	0.07	1.33	3.95E-05
	90897564	rs2736990	G	0.44	1.35	2.88E-08	1.32E-07	0.45	1.24	3.98E-08
17q12–21 (MAPT)	41074926	rs393152	A	0.75	1.32	2.68E-05	1.43E-04	0.76	1.31	2.20E-08
	41279463	rs12185268	A	0.76	1.32	3.44E-05	1.62E-04	0.76	1.30	3.59E-08
	41279910	rs12373139	G	0.76	1.33	1.81E-05	7.60E-05	0.76	1.30	2.77E-08
	41281077	rs17690703	C	0.72	1.34	3.94E-06	6.61E-06	0.72	1.24	1.37E-06
	41347100	rs17563986	A	0.75	1.34	1.30E-05	5.65E-05	0.76	1.31	2.58E-08
	41412603	rs1981997	G	0.76	1.33	2.20E-05	8.81E-05	0.76	1.30	4.61E-08
	41436901	rs8070723	A	0.75	1.33	2.19E-05	8.91E-05	0.76	1.30	2.61E-08
	41544850	rs7225002	A	0.59	1.27	2.72E-05	4.08E-05	0.57	1.23	1.14E-07
	41602941	rs2532274	A	0.75	1.33	2.21E-05	1.06E-04	0.75	1.28	2.92E-07
	41605885	rs2532269	T	0.75	1.33	1.90E-05	8.58E-05	0.76	1.29	1.11E-07
	41648797	rs2668692	G	0.76	1.33	1.97E-05	8.20E-05	0.76	1.29	1.22E-07
4p15 (BST1)	15346446	rs4698412	A	0.52	1.28	6.88E-06	1.96E-06	0.55	1.08	0.0247
Newly identified loci										
1p36.22	11880226	rs12724129	T	0.34	1.26	4.35E-05	4.45E-05	0.40	0.99	[f]
1p36.11	26448619	rs10902724	A	0.05	1.55	5.13E-05	5.00E-04	0.07	0.96	[f]
2q14.3	126112282	rs1365783	G	0.42	1.26	4.33E-05	9.10E-05	0.46	0.94	[f]
2q21.3	135011650	rs621341	T	0.28	1.30	5.11E-06	4.37E-04	0.48	1.08	0.0277
	135196450	rs6723108	G	0.31	1.25	6.85E-05	3.20E-03	0.51	1.11	0.0053
	135318626	rs6729702	G	0.46	1.26	1.75E-05	6.02E-04	0.64	1.04	0.15
	135339278	rs6430552	C	0.46	1.26	1.89E-05	6.95E-04	0.64	1.05	0.14
	135367236	rs6714498	T	0.46	1.26	1.84E-05	6.20E-04	0.64	1.05	0.14
2q22	136611978	rs4954564	A	0.52	1.27	1.21E-05	4.07E-04	0.74	1.03	0.27
	136722668	rs6430612	T	0.41	1.27	1.35E-05	1.89E-03	0.65	1.03	0.23
	136730076	rs10221893	T	0.41	1.27	1.31E-05	1.88E-03	0.65	1.03	0.22
2q35	216463846	rs6741233	C	0.87	1.51	9.34E-06	4.46E-04	0.93	1.01	0.46
4p16	11054284	rs368039	A	0.11	1.41	1.52E-05	2.11E-05	0.13	0.86	[f]
5p15.2	10016889	rs1428954	G	0.53	1.27	1.65E-05	3.73E-04	0.57	1.04	0.18
	10026935	rs10072891	G	0.53	1.27	2.02E-05	6.68E-04	0.57	1.04	0.16
	10037418	rs38065	A	0.65	1.29	1.50E-05	2.19E-04	0.69	1.02	0.33
5q12.1	60896208	rs1423326	T	0.60	1.27	2.10E-05	6.54E-04	0.65	1.02	0.36
5q22.2	112814742	rs26990	C	0.13	1.41	6.67E-06	8.67E-06	0.19	0.94	[f]
6q12	70963882	rs9360414	T	0.38	1.25	5.34E-05	4.00E-05	0.38	1.05	0.13
10p14	6933911	rs10905042	C	0.06	1.53	3.08E-05	3.82E-04	0.07	1.02	0.42
11p12	36589978	rs12419750	A	0.89	1.47	4.96E-05	1.65E-05	0.90	0.97	[f]
	36600652	rs1391542	A	0.89	1.47	5.44E-05	1.89E-05	0.90	0.98	[f]
	36613848	rs7128419	A	0.89	1.47	4.91E-05	1.71E-05	0.90	0.98	[f]
	36618299	rs12271660	A	0.90	1.50	5.64E-05	4.56E-05	0.92	0.96	[f]
	36684837	rs12294719	T	0.79	1.44	5.42E-07	6.72E-07	0.81	1.00	0.48
	36687460	rs1533588	A	0.79	1.41	1.79E-06	3.78E-06	0.82	0.97	[f]
11q13.5	75709727	rs12295401	T	0.06	1.53	4.40E-05	5.16E-05	0.06	1.04	0.29
12p13.3	760163	rs11064524	G	0.20	1.32	2.80E-05	1.21E-04	0.24	1.08	0.0447
12q21.31	82691472	rs7954761	A	0.60	1.34	2.09E-07	2.59E-07	0.61	0.99	[f]
12q24	105474117	rs4964469	A	0.33	1.27	2.73E-05	1.30E-04	0.37	1.11	0.0045
	105513235	rs1035767	T	0.11	1.42	1.50E-05	2.15E-05	0.11	0.98	[f]
13q34	113253980	rs2259599	G	0.83	1.39	3.74E-05	5.61E-03	0.88	0.96	[f]
17p13.2	4376339	rs9899558	G	0.73	1.34	5.04E-06	3.82E-04	0.77	0.98	[f]
22q11.23	22917303	rs9608247	A	0.16	1.33	2.99E-05	9.67E-05	0.17	0.95	[f]

[a]Risk allele in stage-1 data.
[b]Risk allele frequency in controls.
[c]Logistic tests corrected for genomic inflation.
[d]Logistic tests including 2PCs as covariates.
[e]*P*-values shown when the direction of effect in stage-1 and stage-2 data are consistent.
[f]One-tailed $P > 0.5$.

($P = 0.018$). Of the four regions, only one (12p13.3) showed no evidence of association from the combined analysis of the two replication datasets. Overall, evidence of association was consistently stronger with the region of the newly identified PD risk locus than with BST1, in each replication sample as well as in the combined (genome-wide and two replication samples) data (12q24, $P = 2.38 \times 10^{-7}$; BST1, $P = 1.79 \times 10^{-6}$).

Table 2. GWAS and replication: loci considered to follow-up

Locus	Stage *n* (case/control) Position (bp)	SNP	Stage-1 (France) (1039/1984) RA[a]	RAF[b]	%OR (95% CI)	*P* (two-tailed)	Stage-2 (UK) (1705/5200) RAF	OR (95% CI)	*P* (one-tailed)[c]	Stage-3 (France/Australia) (1527/1864) RAF	OR (95% CI)	*P* (one-tailed)[c,d]	Combined Stage 2 + 3 (3232/7064) OR (95% CI)	*P* (one-tailed)[c,d]	Stage 1 + 2 + 3 (4271/9048) OR (95% CI)	*P* (two-tailed)[d]
BST1	15346446	rs4698412	A	0.52	1.28 (1.15–1.42)	6.9E-06	0.55	1.08 (1.00–1.17)	0.0247	0.54	1.10 (1.00–1.21)	0.0290	1.09 (1.02–1.16)	0.00333	1.14 (1.08–1.20)	1.79E-06
2q21.3	135011650	rs6721341	T	0.28	1.30 (1.16–1.46)	5.1E-06	0.48	1.08 (1.00–1.17)	0.0277	0.35	1.02 (0.92–1.13)	0.3399	1.06 (1.00–1.13)	0.03500	1.12 (1.04–1.18)	7.50E-05
	135196450	rs6723108	G	0.31	1.25 (1.12–1.40)	6.9E-05	0.51	1.11 (1.02–1.20)	0.0053	0.38	1.03 (0.93–1.14)	0.3047	1.08 (1.01–1.15)	0.00916	1.12 (1.06–1.18)	2.99E-05
12p13.3	760163	rs11064524	G	0.20	1.32 (1.16–1.50)	2.8E-05	0.24	1.08 (0.99–1.18)	0.0447	0.25	0.93 (0.83–1.04)	[e]	–	–	–	–
12q24	105474117	rs4964469	A	0.33	1.27 (1.13–1.41)	2.7E-05	0.37	1.11 (1.03–1.21)	0.0045	0.34	1.12 (1.01–1.24)	0.0175	1.12 (1.05–1.19)	0.00036	1.16 (1.09–1.22)	2.38E-07

[a]Risk allele in stage-1 data.
[b]Risk allele frequency in controls; %odds ratio computed for the stage-1 risk allele.
[c]*P*-values shown when the direction of effect in stage-1 and each replication data are consistent.
[d]*P*-values from stratified association tests.
[e]One-tailed $P > 0.5$.

The evidence of association ($P_{GC} < 1.35 \times 10^{-5}$, Tables 1 and 3) that we detected with several SNPs in the 3′ block of linkage disequilibrium (LD) of the SNCA locus (Fig. 2A), including the two SNPs reaching genome-wide significance in our scan sample, is highly consistent with previous PD GWAS studies (5,6).

MAPT is located in a large block of LD on chromosome 17q12–q22, which contains several additional genes (Fig. 2B). Previous studies (9,10) have identified a large haplotypic block associated with PD, with H1 and H2 being the at-risk and the protective haplotype, respectively. Our two most associated SNPs in the 17q12–q22 region are located within this haplotypic block: rs17690703 ($P_{GC} = 3.9 \times 10^{-6}$) and rs17563986 ($P_{GC} = 1.3 \times 10^{-5}$), the latter being at MAPT. In addition, H2 is tagged by the minor alleles of four of our genotyped SNPs: rs12185268/G, rs12373139/A, rs1981997/A and rs8070723/G. In our scan data, we found the same H1/H2 association signals, with all minor alleles of these four SNPs being significantly associated ($P_{GC} < 3.44 \times 10^{-5}$) with a decreased risk of PD (Table 1).

The BST1 gene has previously been associated with PD in a Japanese GWAS at a genome-wide significance level (5). Strong evidence of association for rs4698412 was found in the Japanese scan ($P = 5.3 \times 10^{-5}$, OR = 1.25) and in the combined (scan + replication) data ($P = 1.8 \times 10^{-8}$, OR = 1.24) (5). A much weaker signal was obtained in the US/UK/German data, in both the scan ($P = 0.09$, OR = 1.07) and the combined ($P = 0.03$, OR = 1.06) data (6). Here, we report strong evidence of association of PD with BST1 (combined $P = 1.79 \times 10^{-6}$, OR = 1.14). The most associated SNP (rs4698412) maps to a 15 kb LD-block (Fig. 2C) and is in high LD ($r^2 = 0.74/0.79$) with the next top two BST1 variants (Table 3). Despite the variation in the allele frequency of the risk allele between the Japanese (RAF = 0.33) and the European (RAF = 0.52–0.56) samples (Tables 2 and 3), we found marked homogeneity in the direction of effects across the groups, but effect sizes seemed to be lower in European than in Japanese samples. BST1 has been proposed to play a role in generating cyclic ADP-ribose that serves as a second messenger for Ca^{2+} mobilization in endoplasmic reticulum and thus Ca homeostasis-related BST1 could be a cause of selective vulnerability of dopaminergic neurons in PD (11).

Our most associated SNP, on 12q24 (combined $P = 2.38 \times 10^{-7}$, OR = 1.16), is 26 kb centromeric of RFX4 (Regulatory factor X4) (Fig. 2D). Two other close genes, POLR3B (Polymerase RNA III polypeptide B) and RIC8B (Resistance to inhibitors of cholinesterase 8 homolog B), are 200 kb centromeric and telomeric of the 12q24 SNP, respectively. The RFX proteins belong to the winged-helix subfamily of helix–turn–helix transcription factors. The *RFX4_v3* transcript variant is the only *RFX4* isoform that is significantly expressed in the fetal and adult brain, and its expression is restricted to the brain. In addition, it has a role in the transcription of many genes involved in brain morphogenesis, such as the signaling components in the wnt, bone morphogenetic protein (BMP) and retinoic acid (RA) pathways. In particular, cx3cl1, a CX3C-type chemokine gene, which is highly expressed in brain in response to injury or infection and regulates intracellular calcium concentration, was downregulated

Table 3. Association results of previously reported PD loci

		Japanese samples							European samples										
		Scan phase					Combined (scan + replication)		Scan phase					Combined (scan + replication)		France-GWAS			
SNP	bp	GWAS	Risk all	RAF	OR	P^a	OR	*P*	GWAS	Risk all	RAF	OR	*P*	OR	*P*	Risk all	RAF	OR	P_{GC}
(A) Genome-wide significant loci																			
4q22 (SNCA)																			
rs11931074	90858538	S1	T	0.58	1.50	*6.2E-13*	1.37	*7.4E-17*	S2	T	0.07	1.49	*4.8E-08*	1.46	1.6E-14	T	0.07	1.52	1.3E-05
rs356220	90860363	NA							NA							T	0.35	1.37	*2.82E-08*
rs3857059	90894261	S1	G	0.59	1.49	*1.2E-12*	1.36	*5.7E-16*	S2	G	0.07	1.49	*3.6E-08*	1.48	3.7E-15	G	0.07	1.54	1.0E-05
rs2736990	90897564	NA							S2	C	0.46	1.27	*5.7E-09*	1.23	2.2E-16	G	0.44	1.35	*2.9E-08*
vrs6532194	90999925	S1	T	0.6	1.44	*7.0E-11*	1.32	*4.2E-13*	NA							T	0.09	1.22	0.028
17q12-q22 (MAPT)																			
rs393152	41074926	NA							S2	A	0.78	1.32	1.4E-07	1.30	2.0E-16	A	0.75	1.32	2.7E-05
rs17563986	41347100	NA							S2	T	0.78	1.30	3.4E-07	1.28	1.7E-14	A	0.75	1.34	1.3E-05
rs199533	42184098	NA							S2	C	0.80	1.33	*5.1E-08*	1.28	1.1E-14	G	0.78	1.28	3.0E-04
4p15 (BST1)																			
rs3213710	15326419	NA							NA							A	0.50	1.24	7.7E-05
rs4698412	15346446	S1	A	0.33	1.25	5.3E-05	1.24	*1.8E-08*	S2	A	0.56	1.07	0.09	1.06	0.03	A	0.52	1.28	6.9E-06
rs4538475	15347035	S1	A	0.36	1.25	4.1E-05	1.24	*3.9E-09*	NA							A	0.83	1.15	0.07
rs12646913	15348374	NA							S2	A	0.92	1.18	0.04	1.09	0.03	A	0.91	1.19	0.08
rs4698120	15352430	NA							NA							C	0.53	1.24	7.4E-05
1q32 (PARK16)																			
rs16856139	203905087	S1	C	0.86	1.50	2.6E-06	1.46	*1.0E-07*	NA							T	0.06	1.10	-
rs947211	204019288	S1	G	0.52	1.23	1.2E-04	1.3	*1.5E-12*	NA							G	0.78	1.10	0.14
rs823156	204031263	S1	A	0.83	1.40	1.2E-05	1.37	*3.6E-09*	S2	T	0.82	1.12	4.3E-03	1.12	7.6E-04	A	0.81	1.17	0.03
rs708730	204044403	S1	A	0.82	1.37	2.6E-05	1.33	*2.4E-08*	NA							A	0.82	1.12	0.13
12q12 (LRRK2)																			
rs11564162	38729159	NA							S2	T	0.81	1.28	4.0E-05	1.15	9.5E-05	G	0.17	1.11	-
rs2708453	38764919	S1	T	0.08	1.41	7.5E-05	1.38	9.7E-08	NA							T	0.16	1.16	0.04
rs2896905	38779683	NA							S2	C	0.60	1.22	5.0E-06	1.07	7.8E-03	G	0.64	1.02	0.75
rs1491923	38877384	NA							S2	C	0.31	1.20	2.2E-04	1.14	1.6E-05	G	0.31	1.10	0.09
(B) Suggestive loci																			
4p16 (GAK)																			
rs1564282	842 313	NA							S3	T	0.09	1.7	6.0E-06			T	0.09	1.27	0.008
rs11248060	954 359	NA							S3	T	0.10	1.69	3.4E-06			T	0.11	1.18	0.05

S1: Table 1 from Satake *et al.*, *Nat. Genet.*, 2009; Japanese data – scan phase (1078 PD/2521 controls). S2: Table 2 from Simon-Sanchez *et al.*, *Nat. Genet.*, 2009; US/GE/UK data – scan phase (1745 PD/4047 controls). S3: Table 2 from Pankratz *et al.*, *Hum. Genet.*, 2009; US (PROGENI + GenePD) data – (857 familial PD cases/867 controls).
[a]Genome-wide significant values ($<10^{-7}$) are italicized.

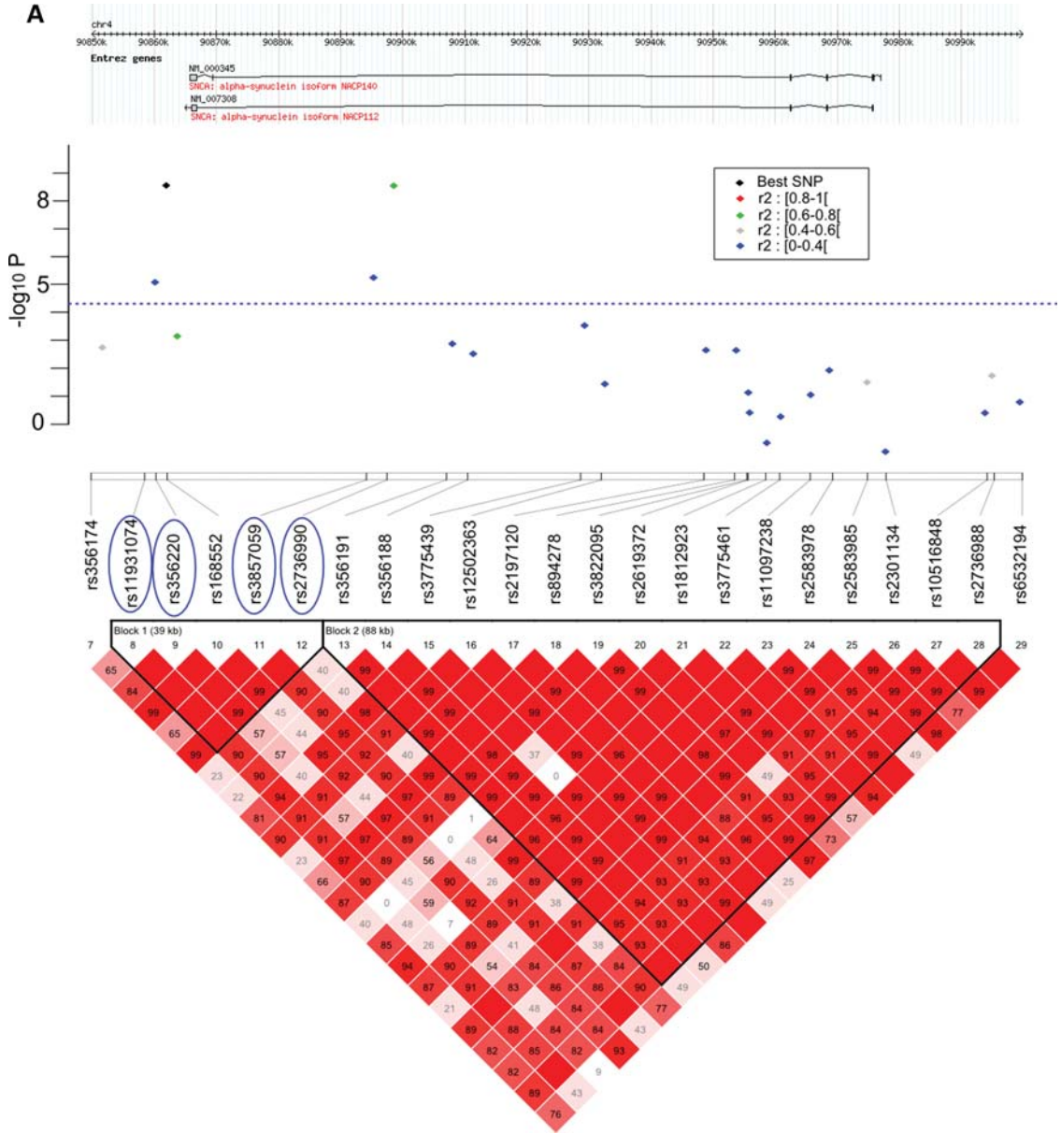


Figure 2. Regional association plots and LD structure for the four PD risk loci (**A**) 4q22/SNCA, (**B**) 17q12–q22/MAPT, (**C**) 4p15/BST1 and (**D**) 12q24/RFX4. The –log10 *P*-values (logistic regression tests corrected for genomic inflation) in the GWAS stage. In each panel, the blue horizontal line indicates a *P*-value of 5×10^{-5}. Pairwise linkage disequilibrium (D') values are displayed and the SNPs with the strongest association signals are circled. SNPs are color-coded for LD relationships (r^2) to the best (colored in black) SNP: red, $0.8 \leq r^2 < 1$; green, $0.6 \leq r^2 < 0.8$; gray, $0.4 \leq r^2 < 0.6$; blue, $0 \leq r^2 < 0.4$. Positions are NCBI build 36 coordinates. Intron and exon structures of genes are taken from the UCSC Genome Browser.

in RFX4_v3-null mice (12). This allows speculation that RFX4 and BST1 are functionally linked and indirectly involved in the regulation of intracellular Ca^{2+} concentrations, which plays an important role in various cellular functions and cell death. Finally, polymorphisms in RFX4 have been shown to be risk factors for the bipolar disorder, manic-depressive illness (13). A recent study showed that a substantial proportion (10–15%) of top GWAS hits, so far identified, are e-quantitative trait loci (eQTLs), i.e. associated with gene expression levels (14). We have initiated eQTL analysis using an existing brain expression database (15), but so far failed to identify any association of the PD-associated rs4964469 SNP with the expression of known genes contained within the 12q24 region.

In conclusion, we have conducted a large GWAS of PD in three case–control samples from France, the UK and Australia. The GWAS stage has 75% and 33% power to detect the loci of the effect sizes observed in stage-1 data for the 12q24 variant (OR = 1.27) at a significance of $P < 5 \times 10^{-5}$ and $P < 10^{-7}$, respectively. In the scan-step, we detected genome-wide significance of association with PD for two SNPs on 4q22, and strong evidence of association with 17q12–q22 SNPs. The two regions encompass previously reported loci: SNCA and MAPT, respectively.

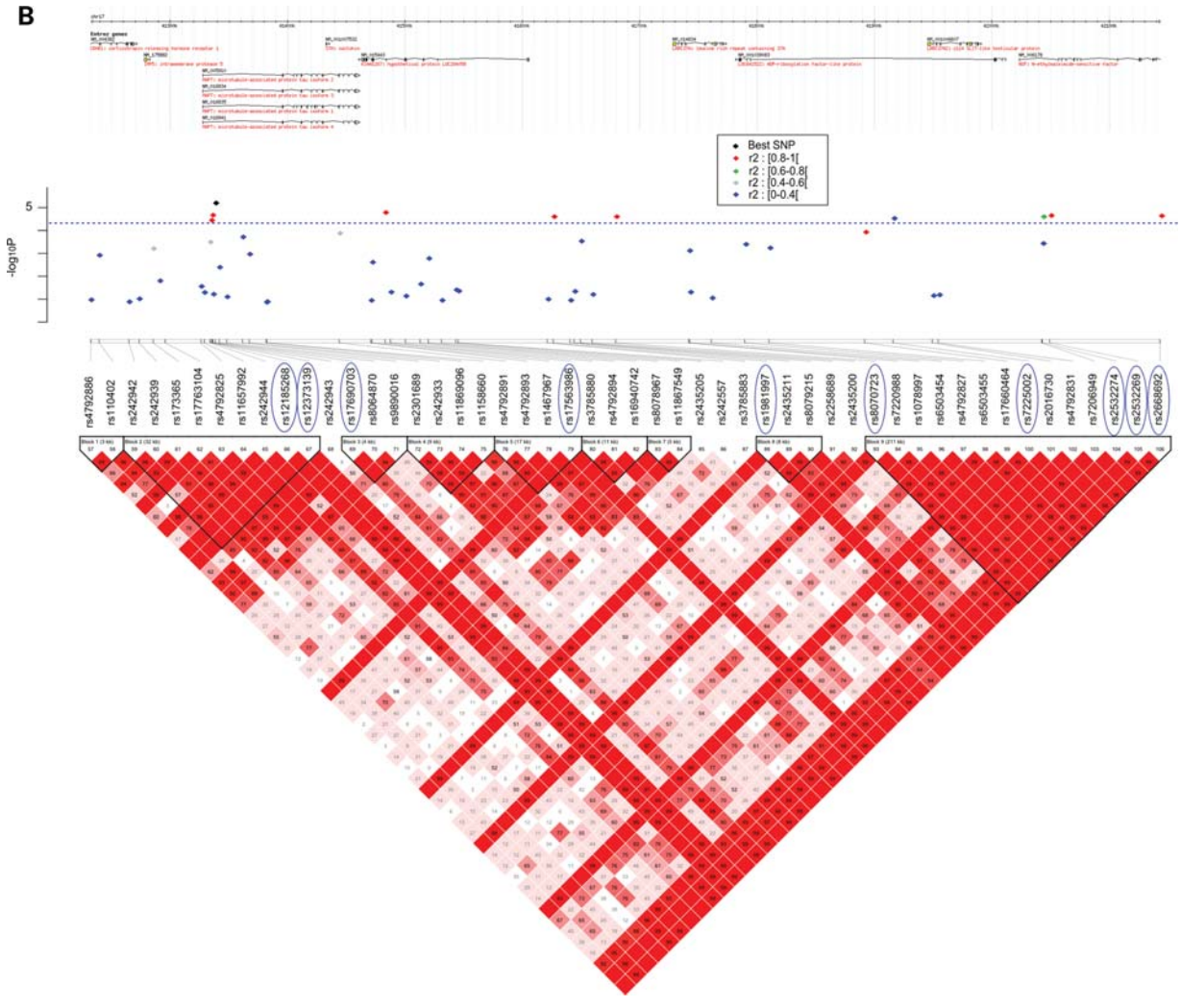


Figure 2. Continued

In addition, we confirmed, for the first time in subjects of European ancestry, the association of PD with 4p15/BST1, recently identified in Japanese samples. Finally, we identified a new locus on 12q24, potentially associated with PD. Further replication studies conducted in large case–control samples are warranted to evaluate the contribution of this locus to PD risk.

MATERIALS AND METHODS

Sample ascertainment and diagnostic criteria

The main characteristics of the three case–control samples are shown in Table 4.

Stage-1 subjects. The total number of cases and controls from France included in stage 1 was 1070 and 2023 controls, respectively.

- **PD subjects:** Patients were recruited through the French network for the study of Parkinson's disease Genetics (PDG) that comprises 15 university hospitals across France. Definite and probable PD was defined according to standard criteria. Definite PD required at least two of three cardinal signs (akinesia and/or rigidity and/or tremor) and absence of exclusion criteria (ophthalmoplegia, pyramidal or cerebellar signs, early dementia, urinary incontinence or postural instability and prior exposure to neuroleptic drugs), and a positive and sustained response to levodopa therapy. Probable PD required at least two of the five following criteria: the parkinsonian triad, a good response to levodopa therapy and asymmetrical onset. Most (>80%) of the PD cases fulfilled the criteria for definite PD. Patients were selected in an effort to enrich for individuals who may have greater genetic predisposition to PD, through selection of cases with a positive family history of PD (Table 4). Cases were of European origin, mostly French ($n = 930$). Subjects diagnosed genetically with known *PARK* mutations (*SNCA*, *LRRK2*, *parkin* and *PINK1*) were excluded.

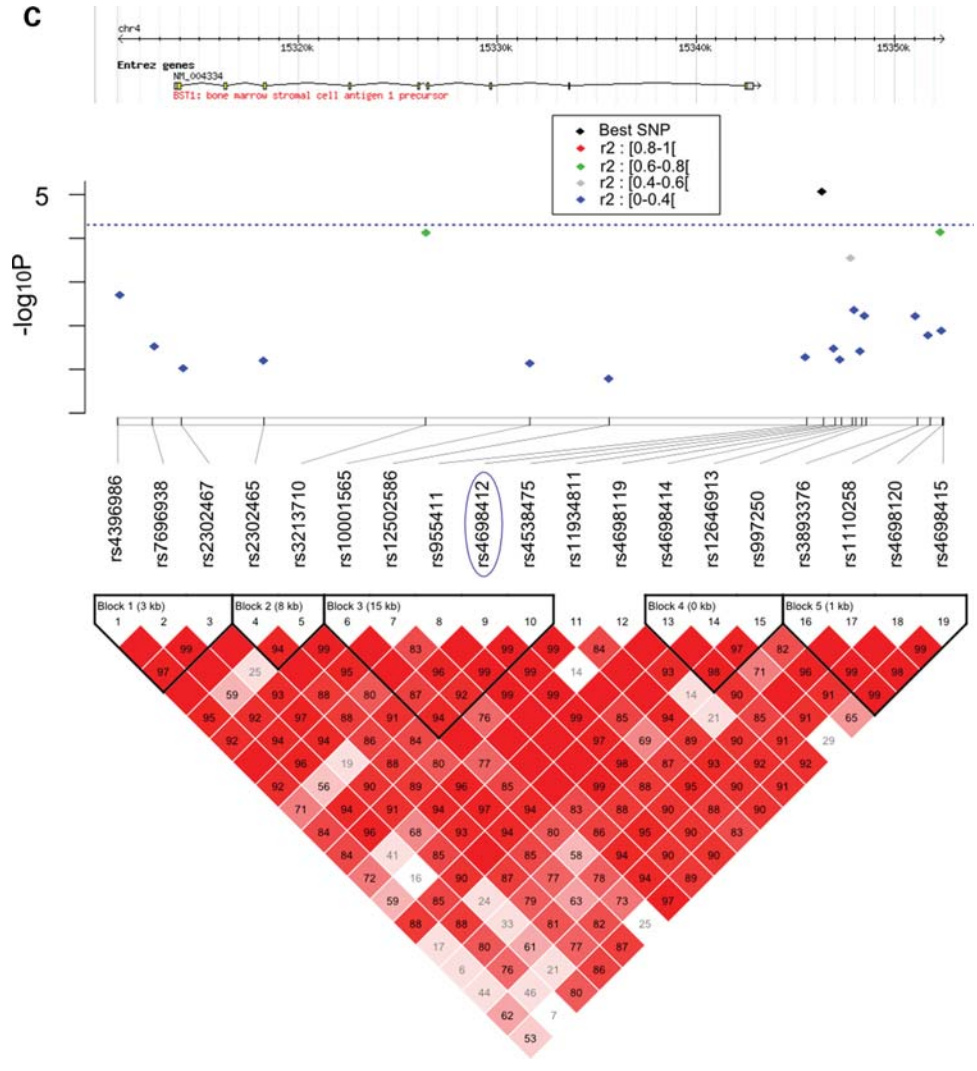


Figure 2. Continued

- 3C neurologically normal controls: The French Three-City (3C) cohort is a population-based, prospective (4-year follow-up) study of the relationship between vascular factors and dementia, carried out in three French cities: Bordeaux (Southwest France), Dijon (central eastern France) and Montpellier (Southeast France) (16). Participants (>9000) are non-institutionalized subjects, over 65 years of age, randomly selected from the electoral rolls of each city. Patients with Alzheimer's disease or other types of dementia, and individuals for whom information on their dementia status during the 4-year follow-up was missing were further excluded. Here, we used a sample of 2023 neurologically normal subjects matched on gender with PD cases, randomly selected from all the participants.

Stage-2 subjects. *In silico* replication sample: we exchanged genome-wide association data with the WTCCC2 PD study group (Spencer *et al.*, submitted). This case–control study consisted of 1705 PD cases and 5200 controls from the 1958 Birth Cohort and from the OK Blood Services Controls (17).

Stage-3 subjects. *De novo* genotyping was conducted in two independent case–control datasets from France (872 PD, 1440 controls) and Australia (655 PD, 424 controls). The subjects from France were combined from three French studies: TERRE (207 cases, 468 controls), PARTAGE (313 cases, 593 controls) and an extension of PDG (352 cases, 378 controls). The extension PDG study includes patients who were not available at the time of the stage-1 genotyping execution and neurologically normal spouses of PDG patients. In cases, the mean age at examination and the mean age of onset of PD is 59 (30–86) and 50 (20–84) years, respectively. The mean age of controls is 60 (31–85) years. In PARTAGE, patients and controls were identified among affiliates of the Mutualité Sociale Agricole (MSA) from five French districts. Parkinsonism was defined as the presence of at least two cardinal signs (rest tremor, bradykinesia, rigidity, impaired

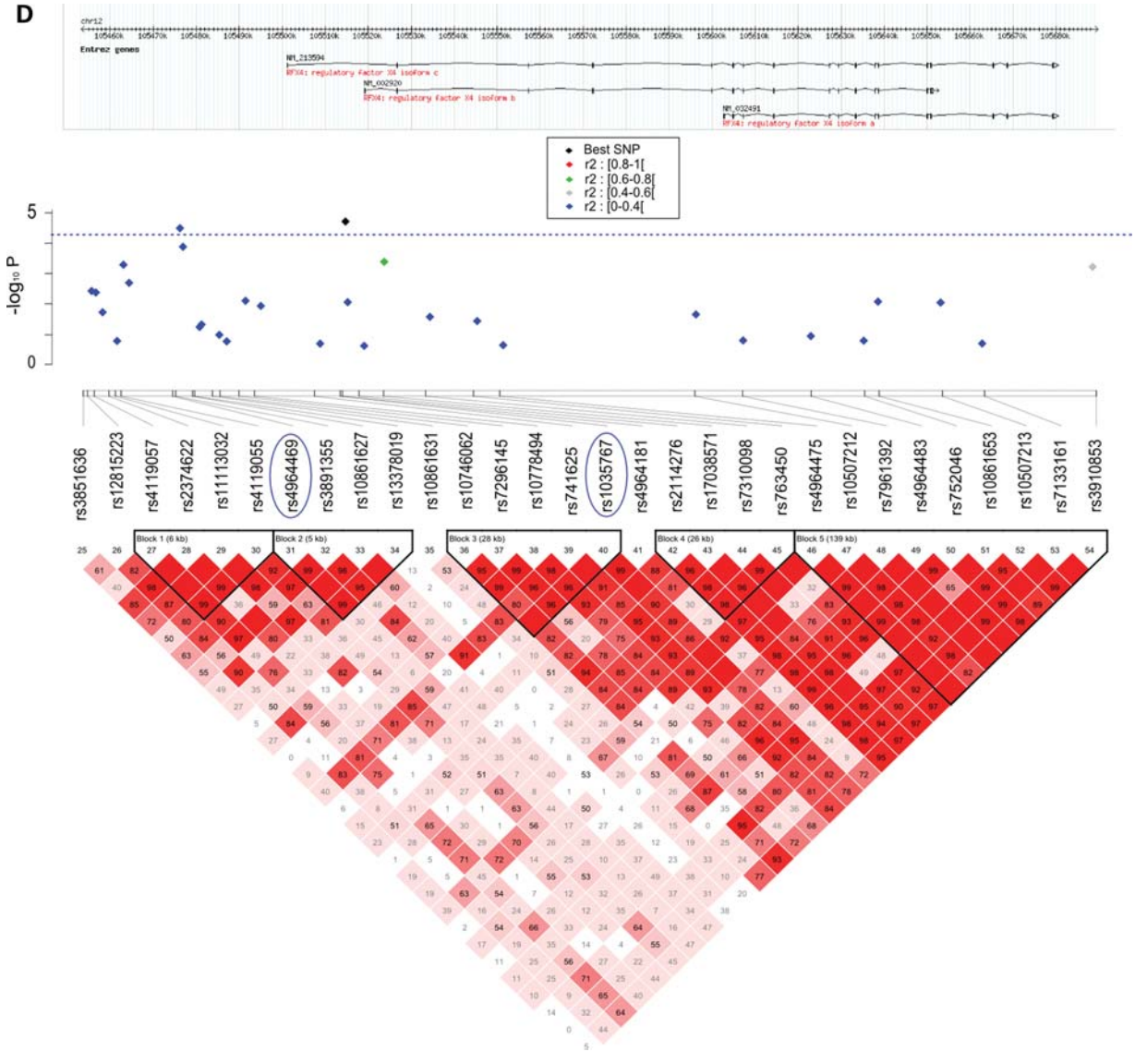


Figure 2. Continued

postural reflexes); PD was defined as the presence of parkinsonism after exclusion of other causes of parkinsonism. Controls were randomly selected from all MSA affiliates in the same districts and matched for sex and age (±2 years). DNA was collected from saliva (Oragene kit). Cases and controls have a mean age of 67 (37–79) years, and the mean age of onset of disease is 63 (35–75) years. TERRE is based on a similar protocol (18), but DNA was collected from blood; the mean age in cases and controls is 73 (46–82) years, and the mean age of onset is 66 (39–80) years in cases.

Australian study: Subjects with PD were recruited from one private and two public movement disorder clinics in Brisbane. Controls were electoral roll volunteers and patient spouses, excluding the subjects demonstrating signs of parkinsonism (19). The mean age is 72 (34–105) and 74 (33–107) years in controls and cases, respectively; the mean age of onset is 59 (23–96) years in cases. Only Caucasian subjects were included in stage 3; in the Australian study, analyses were restricted to participants who reported having four European grandparents (>85% British). There was no overlap between the subjects used in the replication datasets and those included in the stage-1 data. Written informed consent was obtained for all participating subjects and research protocols were approved by local ethics committees.

Genotyping

Stage-1 genotyping. DNA samples of PDG cases and 3C controls were transferred to the French Centre National de Génotypage. First-stage samples that passed DNA quality control (QC) (1064 PD cases and 2023 controls) were genotyped with Illumina Human610-Quad BeadChip and subjected to standard QC procedures.

Table 4. Samples used (post-QC) in this study

Center Genotyping platform	Stage-1 Scan French Illumina 610-Quad	Stage-2 Replication UK Illumina 650Y	Stage-3 Replication French–Australian Illumina GoldenGate	Total
Cases	1039	1705	1527	4271
Sex ratio: M/F	1.42	1.37	1.42	
Age: mean ± SD (*n*)[a]	57.5 ± 16.6 (1003)	NA	69.0 ± 12.7 (1365)	
AOO: mean ± SD (*n*)[a]	48.9 ± 12.8 (970)	65.2 ± 11.3 (1109)	61.3 ± 12.4 (1351)	
FH+ (%)	47	0	17	
Controls	1984	5200	1864	9048
Sex ratio (M/F)	1.33	1.02	1.05	
Age: mean ± SD (*n*)[a]	73.7 ± 5.4 (1984)	51	68.1 ± 10.0	
Total	3023	6905	3391	13 319

[a]Number of subjects for which age/age of onset of disease is known.

Stage-2 genotyping. This WTCCC2 PD study sample was genotyped by the Welcome Trust Case–Control Consortium using the Illumina 650Y genotyping array (Spencer *et al.*, submitted).

Stage-3 genotyping. Genotyping in the extended PDG sample was carried out in the UMR/S 975 laboratory, using predesigned TaqMan probes (C_537709_10/ rs621341; C_29330880_10/ rs6723108; C_12096605_10/ rs11064524; C_2775670_10/ rs4964469; C_1216796_10/ rs4698412) on an ABI 7500 Real-Time PCR system Applied Biosystems, Foster City, CA, USA), according to the manufacturer's instructions. Data were then analyzed using the 7500 software v.2.0.1. The TERRE/PARTAGE and Australian samples were genotyped using the Sequenom MassARRAY platform, with the iPLEX protocol (Genoscreen, France). The basic protocol involves a multiplex primer extension followed by matrix-assisted laser desorption ionization-time of flight mass spectroscopy detection. In order to avoid any genotyping bias, cases and controls were randomly mixed when genotyping and, laboratory personnel were blinded to case–control status.

Quality control of France GWAS scan data

Various stringent QC filters were applied to remove poorly performing SNPs and samples using tools implemented in PLINK version 1.7 (20).

SNP QC: Markers were removed if they had a genotype-missing rate >0.03 or a minor allele frequency (MAF) < 0.05 or a Hardy–Weinberg $P \leq 10^{-5}$. This SNP QC step led to the removal of 74 660 autosomal SNPs. Thus, subsequent analyses were based on 492 929 SNPs.

Individual QC: Samples were removed based on standard exclusion criteria: call rate of $<96\%$ (22 subjects), inconsistencies between reported gender and genotype-determined gender (11 subjects) and genetic relatedness (identity-by-descent estimate >0.14; 6 subjects). Applying these QC filters led to the removal of 39 subjects (14 cases, 25 controls).

Population stratification and principal component analysis: To detect individuals of non-European ancestry, we thinned the SNPs to reduce LD to a set of 55 193 SNPs. To this end, we removed SNPs from the extensive regions of LD (CHR2, CHR5, CHR6, CHR8, CHR11) (21), and excluded SNPs if any pair within a 1000-SNP window had $r^2 > 0.2$. Our stage-1 genotype data were then merged with genotypes at the same SNPs from 381 unrelated European (CEU), Yoruban (YRI) and Asian (CHB and JPT) samples from the HapMap project. Principal component analysis was applied using EIGENSTRAT (22). The two PCs clearly separated the HapMap data into three distinct clusters according to ancestry, and most of our stage-1 samples were clustered with the HapMap European samples (Fig. 3). Thirty-two samples appeared to be ethnic outliers (including one subject clearly sharing African ancestry) from the European cluster and were excluded from further analysis. The final post-QC scan sample comprised 1039 PD cases and 1984 controls.

Statistical analysis

Association analysis of the genotype data was conducted with PLINK (20).

Stage-1 association analyses. Logistic regression was used to study the allelic association between each SNP and PD assuming an additive genetic model. Our analysis was based on 492 929 SNPs, and on a conservative genome-wide significance threshold of $0.05/492\,929 = 10^{-7}$. The distribution of the association results was found to be marginally inflated (median $\chi^2 = 0.521$); genomic inflation factor $\lambda = 1.14$ ($\lambda_{1000} = 1.10$). Logistic regression analysis adjusted for the two first PCs of the EIGENSTRAT analysis revealed a genomic inflation of 1.03 (median $\chi^2 = 0.472$). As for our primary analyses, we applied the genomic inflation correction method (23); the median of the GC-corrected χ^2 value was 0.447.

Sensitivity analyses: Two further analyses were conducted to assist in the interpretation of results of the identified GWAS SNPs. We performed age-adjusted regression analysis and conducted subgroup analyses of two subtypes of cases against all controls. Cases with a disease onset before 50 years ($n = 428$) were classified as 'early AOO', and cases

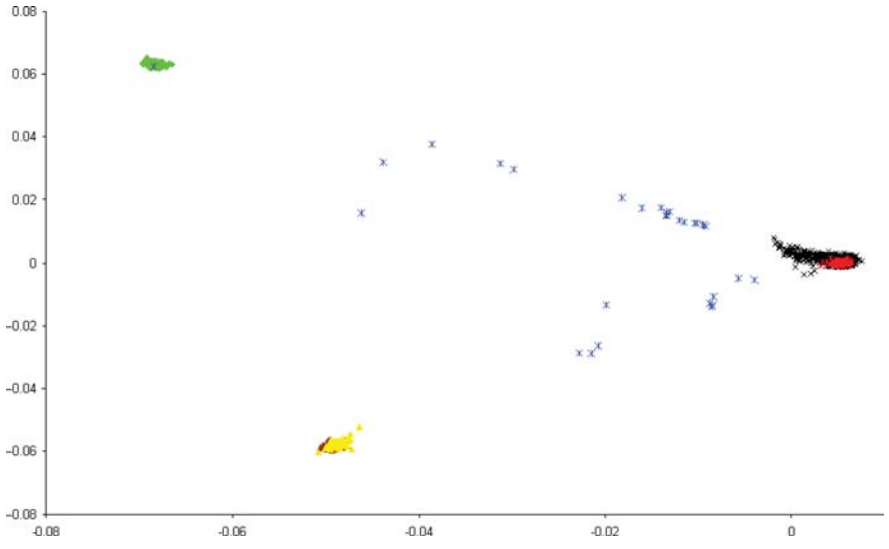


Figure 3. Principal components for our genome-wide stage 1. Plot of the first two principal components from the analysis of our stage-1 (post-QC) data combined with HapMap data. Ethnicity of HapMap samples indicated by color: Africa (YRI) in green, Japan (JPT) in brown, Chinese (6) in yellow and Europe (CEU) in red. Study samples identified to be non-European or not clustering with European samples (outliers) are colored in blue and the remaining study samples assumed to be of European origin are colored in black.

having at least one first-degree relative with PD ($n = 452$) were classified as 'FH+'.

Stage-2 in silico *association analyses*. Statistical data (ORs, effective sample sizes and nominal *P*-values for each of the 50 top SNPs) in the UK sample were obtained from the WTCC2 PD study group that used similar analytical methods (Spencer *et al.*, submitted).

Stage-3 association analyses. For the *de novo* replication stage, we computed association statistics with the Mantel–Haenszel test to control for the potential confounding owing to the geographical center (France versus Australia) for the five SNPs replicated at stage-2. Using raw genotypes from all the study samples, we computed similar stratified (France versus Australia versus UK) association statistics in the combined (stage-2 + stage-3 and stage-1 + stage-2 + stage-3) data.

The PAR associated with the detected variants was estimated with the following formula: $PAR = p\ (OR-1)/[p(OR-1) + 1]$, where p is the frequency of the risk allele in controls, and OR is the odds ratio associated with the risk allele.

AUTHOR CONTRIBUTIONS

S.L. supervised DNA sampling; J.C.C., M.V., E.B., F.D., P.P., P.D., F.T., A.D. and A.B. recruited patients; J.C.C., A.D. and AB supervised clinical work; D.Z. and M.L. supervised PD and 3C GWAS genotyping and DNA QC work; S.L. and J.C.L. supervised genotyping of stage-3 samples. A.E., J.C.L., M.A.L., C.T., G.D.M. and P.A.S. contributed to stage-3 replication; M.S., A.S.P. and M.M. executed QC analyses and performed statistical association analyses; A.E., A.B. and M.M. were involved in obtaining funding; M.M. drafted the manuscript and S.L., A.B. and A.E. contributed to the writing of the final version; A.B. and M.M. conceived and oversaw the design and execution of the GWAS.

ACKNOWLEDGEMENTS

The authors are grateful to the patients and their families. They thank the DNA and Cell Bank of UMR_S975 for sample preparation. We thank the members of the 3C consortium: Drs Annick Alpérovitch, Claudine Berr and Jean-Francois Dartigues for giving us the possibility to use part of the 3C cohort. This study makes use of data generated by the Wellcome Trust Case–Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk.

Conflict of Interest statement. The authors declare no competing financial interests.

FUNDING

This work was supported by the French National Agency of Research (ANR-08-MNP-012).

REFERENCES

1. Lesage, S. and Brice, A. (2009) Parkinson's disease: from monogenic forms to genetic susceptibility factors. *Hum. Mol. Genet.*, **18**, R48–R59.
2. Maraganore, D.M., De Andrade, M., Lesnick, T.G., Strain, K.J., Farrer, M.J., Rocca, W.A., Pant, P.V., Frazer, K.A., Cox, D.R. and Ballinger, D.G. (2005) High-resolution whole-genome association study of Parkinson disease. *Am. J. Hum. Genet.*, **77**, 685–693.
3. Fung, H.C., Scholz, S., Matarin, M., Simon-Sanchez, J., Hernandez, D., Britton, A., Gibbs, J.R., Langefeld, C., Stiegert, M.L., Schymick, J. *et al.* (2006) Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol.*, **5**, 911–916.

4. Myers, R.H. (2006) Considerations for genomewide association studies in Parkinson disease. *Am. J. Hum. Genet.*, **78**, 1081–1082.
5. Satake, W., Nakabayashi, Y., Mizuta, I., Hirota, Y., Ito, C., Kubo, M., Kawaguchi, T., Tsunoda, T., Watanabe, M., Takeda, A. *et al.* (2009) Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. *Nat. Genet.*, **41**, 1303–1307.
6. Simon-Sanchez, J., Schulte, C., Bras, J.M., Sharma, M., Gibbs, J.R., Berg, D., Paisan-Ruiz, C., Lichtner, P., Scholz, S.W., Hernandez, D.G. *et al.* (2009) Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat. Genet.*, **41**, 1308–1312.
7. Tian, C., Kosoy, R., Nassir, R., Lee, A., Villoslada, P., Klareskog, L., Hammarstrom, L., Garchon, H.J., Pulver, A.E., Ransom, M. *et al.* (2009) European population genetic substructure: further definition of ancestry informative markers for distinguishing among diverse European ethnic groups. *Mol. Med.*, **15**, 371–383.
8. Pankratz, N., Wilk, J.B., Latourelle, J.C., Destefano, A.L., Halter, C., Pugh, E.W., Doheny, K.F., Gusella, J.F., Nichols, W.C., Foroud, T. *et al.* (2009) Genomewide association study for susceptibility genes contributing to familial Parkinson disease. *Hum. Genet.*, **124**, 593–605.
9. Healy, D.G., Abou-Sleiman, P.M., Lees, A.J., Casas, J.P., Quinn, N., Bhatia, K., Hingorani, A.D. and Wood, N.W. (2004) Tau gene and Parkinson's disease: a case-control study and meta-analysis. *J. Neurol. Neurosurg. Psychiatry*, **75**, 962–965.
10. Zhang, J., Song, Y., Chen, H. and Fan, D. (2005) The tau gene haplotype h1 confers a susceptibility to Parkinson's disease. *Eur. Neurol.*, **53**, 15–21.
11. Zhang, D., Stumpo, D.J., Graves, J.P., Degraff, L.M., Grissom, S.F., Collins, J.B., Li, L., Zeldin, D.C. and Blackshear, P.J. (2006) Identification of potential target genes for RFX4_v3, a transcription factor critical for brain development. *J. Neurochem.*, **98**, 860–875.
12. Chan, C.S., Gertler, T.S. and Surmeier, D.J. (2009) Calcium homeostasis, selective vulnerability and Parkinson's disease. *Trends Neurosci.*, **32**, 249–256.
13. Glaser, B., Kirov, G., Bray, N.J., Green, E., O'donovan, M.C., Craddock, N. and Owen, M.J. (2005) Identification of a potential bipolar risk haplotype in the gene encoding the winged-helix transcription factor RFX4. *Mol. Psychiatry.*, **10**, 920–927.
14. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. and Lathrop, M. (2009) Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.*, **10**, 184–194.
15. Myers, A.J., Gibbs, J.R., Webster, J.A., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M., Leung, D., Bryden, L., Nath, P. *et al.* (2007) A survey of genetic human cortical gene expression. *Nat. Genet.*, **39**, 1494–1499.
16. Alperovitch, A. (2003) Vascular factors and risk of dementia: design of the Three-City Study and baseline characteristics of the study population. *Neuroepidemiology*, **22**, 316–325.
17. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature*, **447**, 661–678.
18. Elbaz, A., Clavel, J., Rathouz, P.J., Moisan, F., Galanaud, J.P., Delemotte, B., Alperovitch, A. and Tzourio, C. (2009) Professional exposure to pesticides and Parkinson disease. *Ann. Neurol.*, **66**, 494–504.
19. Sutherland, G.T., Halliday, G.M., Silburn, P.A., Mastaglia, F.L., Rowe, D.B., Boyle, R.S., O'sullivan, J.D., Ly, T., Wilton, S.D. and Mellick, G.D. (2009) Do polymorphisms in the familial Parkinsonism genes contribute to risk for sporadic Parkinson's disease? *Mov. Disord.*, **24**, 833–838.
20. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
21. Price, A.L., Weale, M.E., Patterson, N., Myers, S.R., Need, A.C., Shianna, K.V., Ge, D., Rotter, J.I., Torres, E., Taylor, K.D. *et al.* (2008) Long-range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet.*, **83**, 132–135. author reply 135–139.
22. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.
23. Devlin, B. and Roeder, K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

Saad et al. BMC Proceedings 2011, 5(Suppl 9):S33
http://www.biomedcentral.com/1753-6561/5/S9/S33

BMC Proceedings

PROCEEDINGS Open Access

Comparative study of statistical methods for detecting association with rare variants in exome-resequencing data

Mohamad Saad[1,2*], Aude Saint Pierre[1,2], Nora Bohossian[1,2], Matthias Macé[1], Maria Martinez[1,2]

From Genetic Analysis Workshop 17
Boston, MA, USA. 13-16 October 2010

Abstract

Genome-wide association studies for complex traits are based on the common disease/common variant (CDCV) and common disease/rare variant (CDRV) assumptions. Under the CDCV hypothesis, classical genome-wide association studies using single-marker tests are powerful in detecting common susceptibility variants, but under the CDRV hypothesis they are not as powerful. Several methods have been recently proposed to detect association with multiple rare variants collectively in a functional unit such as a gene. In this paper, we compare the relative performance of several of these methods on the Genetic Analysis Workshop 17 data. We evaluate these methods using the unrelated individual and family data sets. Association was tested using 200 replicates for the quantitative trait Q1. Although in these data the power to detect association is often low, our results show that collapsing methods are promising tools. However, we faced the challenge of assessing the proper type I error to validate our power comparisons. We observed that the type I error rate was not well controlled; however, we did not find a general trend specific to each method. Each method can be conservative or nonconservative depending on the studied gene. Our results also suggest that collapsing and the single-locus association approaches may not be affected to the same extent by population stratification. This deserves further investigation.

Background

Classical genome-wide association studies have successfully detected many common genetic variants that are associated with complex traits. It is likely that low-frequency or rare variants are also contributing to genetic risk [1]. The statistical power to detect phenotypic association with such variants is limited because of the small number of observations for any given variant and a more stringent multiple test correction compared to common variants [2]. The simultaneous analysis of rare variants aims to identify accumulations of minor alleles within the same functional unit (e.g., gene).

Several new methods have been recently proposed to tackle the rare variant problem [2-6]. The principal difference between them lies in the way the information on the multiple rare variants is used. Some methods use a subset of variants that satisfy predefined selection criteria, whereas other methods use all variants. The methods also differ in the way in which the cumulative information on minor alleles within a functional unit is coded. Finally, multivariate collapsing approaches have also been proposed. Most of these recent developments have been applied to association analyses in data from unrelated individuals. A new method has been recently developed [4,6] that can be applied to both unrelated individual and family data.

In this paper, we evaluate and compare the power of different collapsing methods for detecting association of multiple rare variants with a quantitative trait. We first focus on the unrelated individuals data and then incorporate some of these approaches within the general framework of the mixed model for association analysis in the family data set of Genetic Analysis Workshop 17 (GAW17). We tried to answer the following questions: Does the use of a subset of rare variants perform better than using all variants? Do

* Correspondence: mohamad.saad@inserm.fr
[1]INSERM UMR1043, CPTP, CHU Purpan, Toulouse, 31024, France
Full list of author information is available at the end of the article

the collapsing approaches perform similarly with unrelated individual and family data sets? The analyses were performed using the GAW17 data with knowledge of the answers [7].

Methods

We studied the quantitative trait Q1 influenced by 39 variants in nine independent genes.

Statistical association analysis of rare variants

We carry out the association test at the gene level. Assume that a gene G contains J_G variants denoted SNP^j, $j = 1, \ldots, J_G$, and that MAF_j is the minor allele frequency of SNP^j. Let $Y = (y_1, \ldots, y_N)$ be the observations of the phenotype Q1 in N unrelated individuals, and let X_{iG} be the vector of genotypes of the SNPs in gene G for individual i. The genotypes are coded 0, 1, or 2, depending on the number of minor alleles.

Let T_{maf} be a selection criterion on minor allele frequency (MAF) values. The association methods we have investigated vary according to a predefined T_{maf} value (i.e., less than 1%, less than 5%, or less than 50%) and on the number of collapsing groups. They are all based on a linear regression modeling the relationship between the trait Y and the SNP data within a gene. We briefly review these methods in this Methods section. More details are given by Dering et al. [8].

Association testing in the unrelated individuals data set: univariate collapsing approaches

The univariate collapsing approaches use only a subset of variants that satisfy the constraint MAF $\le T_{\text{maf}}$, where T_{maf} is a predefined selection value.

The first univariate collapsing approach is the collapsing and summation test (CAST). Let $X_{iG}(\text{maf})$ be the vector of genotype scores of the SNPs with MAF $< T_{\text{maf}}$, and let $J_G(\text{maf})$ be the length of the vector $X_{iG}(\text{maf})$. The variable $C = C_{iG}(\text{maf})$ $(i = 1, \ldots, N)$ denotes the two collapsing strategies that we used: collapsing absence/presence (CA) and collapsing proportion (CP). For the CA strategy:

$$C_{iG}(\text{maf}) = \begin{cases} 0 & \text{if } \sum_{j=1}^{J_G(\text{maf})} X_{ij}(\text{maf}) = 0, \\ 1 & \text{otherwise,} \end{cases} \quad (1)$$

and for the CP strategy:

$$C_{ij}(\text{maf}) = \sum_{j=1}^{J_G(\text{maf})} \frac{X_{ij}(\text{maf})}{J_G(\text{maf})}. \quad (2)$$

Equation (1) is based on the presence or absence of the minor allele at any rare variant in gene G within an individual [3]. Equation (2) is based on the proportion of rare variants with MAF $\le T_{\text{maf}}$ at which an individual i carries at least one copy of the minor allele [5]. The model is $Y = C\beta + \varepsilon$, where $\varepsilon \sim N\left(0, I\sigma_\varepsilon^2\right)$ and σ^2 is the residual variance.

The effect of β can be tested with a likelihood ratio test that follows a chi-square distribution with 1 degree of freedom (df).

The second univariate collapsing method is the variable-threshold (VT) approach [2], which uses the CP approach to collapse rare SNPs with MAF $< T_{\text{maf}}$ but maximizes the statistic according to T_{maf}. All T_{maf} values observed in the gene G are considered. For each T_{maf}, a regression z-score is computed. Let $z_{\max}$ be the maximum z-score across all T_{maf} values. The test of association is based on $z_{\max}$, and its statistical significance is evaluated empirically by permutation.

The last univariate collapsing method is the weighted-sum (WS) approach [2], which is a generalization of the binary trait weighted-sum approach proposed by Madsen and Browning [4] for quantitative traits. Under this approach, $T_{\text{maf}} = 0.5$ (i.e., all variants in a gene G are used). The collapsing variable C for subject i in the WS approach is given by:

$$C_{iG}(\text{maf} = 0.5) = \sum_{j=1}^{J(\text{maf})} X_{jG}\xi_j, \quad (3)$$

where:

$$\xi_j = \frac{1}{\left[\text{MAF}_j(1 - \text{MAF}_j)\right]^{1/2}}. \quad (4)$$

For each gene G, a genetic score is calculated as:

$$Z_G = \sum_{i=1}^{N} C_{iG}. \quad (5)$$

The significance of Z_G is assessed empirically by permutation.

Association testing in the unrelated individuals data set: combined multivariate and collapsing approach

The combined multivariate and collapsing (CMC) method originally proposed by Li and Leal [3] uses a multiple regression model that contains the CA method's collapsing variable of SNPs with MAF $< T_{\text{maf}} = 1\%$ and includes all k remaining SNPs, $X_{j1,\ldots,k}$, individually.

The multivariate model (denoted here as CMC3) is:

$$Y = \beta_0 \text{CA}(0\text{–}1\%) + \sum_{j=1}^{k} \beta_j X_j. \quad (6)$$

Evidence of association ($\exists j$, $\beta_j \neq 0$, $j = 0, \ldots, k$) is assessed with the likelihood ratio test, which follows a chi-square distribution with $(k + 1)$ df.

Using only the SNPs with MAF ≤ 5%, we extended this model in two ways. In both extensions the multivariate model contains the CA collapsing variable of SNPs with MAF < 1%. In the first variation of this model (denoted CMC1), the multivariate model also contains the CA collapsing variable of the other SNPs (i.e., 1% ≤ MAF ≤ 5%). In contrast, in the second extension (denoted CMC2), the other SNPs are included individually in the multivariate model.

The CMC1 model is then written as:

$$Y = \beta_0 \mathrm{CA}(0\text{–}1\%) + \beta_1 \mathrm{CA}(1\text{–}5\%), \quad (7)$$

and the test of association is a likelihood ratio test with 2 df.

The CMC2 model is the same as Eq. (6), where k is the number of SNPs and 1% ≤ MAF ≤ 5%. Evidence of association is assessed with the likelihood ratio test with $(k + 1)$ df.

Association testing in the unrelated individuals data set: single-marker test

For comparison purposes, we also carried out a single-locus association test. For a gene G, association with each SNP was tested using the likelihood ratio test. For each gene G, we obtained J_G likelihood ratio test statistics, each with 1 df. The evidence of association at the gene level was based on the maximum of the J_G likelihood ratio test statistics.

Single-marker (SM) tests were conducted with PLINK, version 1.07 [9]. The R.2.10.1 software was used for all collapsing approaches except the VT and WS approaches. For these two approaches we used the R script (http://genetics.bwh.harvard.edu/vt/dokuwiki/) [2], and we set the number of permutations to 1,000.

Association testing in the family data set

We used the measured genotype (MG) test [10], which is a linear mixed model:

$$Y_i = X_i\beta + e_i, \quad (8)$$

where:

$$e_i \sim N\left(0, 2\Phi_i\sigma_c^2 + I\sigma_\varepsilon^2\right), \quad (9)$$

σ_c^2 and σ_ε^2 are the polygenic and the residual variances, respectively, and Φ_i is the kinship matrix in family i. The SNP data in relatives were collapsed as described under the CA, CP, and WS collapsing approaches. In these three approaches, the test of association is a likelihood ratio test with 1 df. In addition, we also carried out the bivariate CMC1 approach using the likelihood ratio test with 2 df. We could not evaluate the VT approach because it maximizes T_{maf}. We carried out the MG test using the QTDT software (http://www.sph.umich.edu/csg/abecasis/QTDT/).

Type I error rate and power estimates

The empirical distribution of each association approach was evaluated in unrelated individuals and in family data. Type I error and power rates were estimated by testing association of Q1 to each of the seven false causal genes and each of the nine true causal genes, respectively, using the 200 replicates. Type I error and power rates were derived at a nominal level of $\alpha = 5\%$.

In the unrelated individuals data set, we evaluated association with Q1 using 10 approaches: CA1 and CA5 with T_{maf} = 1% and 5%, respectively; CP1 and CP5 with T_{maf} = 1% and 5%, respectively; and VT, WS, CMC1, CMC2, CMC3, and SM. For the WS and VT tests, we used empirical P-values. For all remaining association tests we used tabulated nominal P-values. In each replicate, we tested for association of Q1 with each of the 16 genes using each of the 10 approaches. For each gene and for each association procedure we computed the proportion of replicates having a P-value ≤ α. For the SM approach, we applied a Bonferroni correction to account for the multiple tests; we computed the proportion of replicates such that the lowest P-value out of the J_G SNPs was less than or equal to α/J_G.

In the family data set, we evaluated similarly the following five approaches: CA1 and CA5 with T_{maf} = 1% and 5%, respectively; CP1 and CP5 with T_{maf} = 1% and 5%, respectively; and SM. We also evaluated the WS approach but used the tabulated P-value derived from a chi-square distribution with 1 df.

Results and discussion

The characteristics of the nine causal and seven noncausal genes are shown in Table 1. The total number of SNPs (causal and noncausal) per gene is given along with their distributions by MAF category. The MAF for the causal variants ranges from 0.07% to 16.5% in the 1000 Genomes Project data (for unrelated individuals), and the number of causal variants per gene varies from 1 (*VEGFC*, *VEGFA*) to 11 (*FLT1*). One causal gene (*VEGFC*) has one single SNP, and thus only one association approach (SM) can be applied. For the noncausal genes, the number of SNPs per gene ranges from 6 (*CTSS*) to 83 (*LY75*), and, as for the causal genes, most (>70%) of the SNPs are uncommon (MAF < 5%).

Estimates of Type I error and power rates in the unrelated individuals data set

Table 2 shows the type I error rates estimated at the gene level of each association approach for the unrelated

Table 1 Characteristics of the studied genes

Chromosome	Gene	*K*	MAF (%)	*V*	*K* (*V*) > 5%	5% >*K* (*V*) > 1%	*K* (*V*) < 1%
Causal genes							
1	*ARNT*	18	0.07; 43.11	5	1 (0)	2 (1)	15 (4)
1	*ELAVL4*	10	0.07; 43.11	2	2 (0)	1 (0)	7 (2)
13	*FLT1*	35	0.07; 29.05	1	3 (1)	7 (2)	25 (8)
5	*FLT4*	10	0.07; 2.08	2	0 (0)	2 (0)	8 (2)
14	*HIF1A*	8	0.07; 1.2	4	0 (0)	1 (1)	7 (3)
19	*HIF3A*	21	0.07; 38.52	3	4 (0)	2 (0)	15 (3)
4	*KDR*	16	0.07; 16.5	10	1 (1)	1 (1)	14 (8)
6	*VEGFA*	6	0.07; 2.37	1	0 (0)	1 (0)	5 (1)
4	*VEGFC*	1	0.07; 0.07	1	0 (0)	0 (0)	1 (1)
Noncausal genes							
1	*PTGFR*	16	0.07; 1.69	0	0	3	13
1	*IFI44*	22	0.07; 11.33	0	1	1	20
1	*FAM73A*	10	0.07; 0.5	0	0	0	10
17	*MAPT*	27	0.07; 35.58	0	5	7	15
1	*CTSS*	6	0.07; 33.28	0	1	1	4
5	*FOXI1*	15	0.07; 37.30	0	2	0	12
2	*LY75*	83	0.07; 45.91	0	11	12	60

K, number of variants in gene; *V*, number of true causal variants in gene.

individuals data set. As can be seen, the type I error rate is not well controlled no matter which association approach is used: The rates can be higher or lower than expected. For some genes, almost all association approaches show inflated type I error rates (e.g., *MAPT*, *IFI44*). Conversely, for some other genes (*FOXI1*, *LY75*), the type I error rates of some approaches are inflated, whereas the other approaches tend to be conservative.

Overall, the SM and CMC3 approaches appear to have inflated type I errors more frequently. Interestingly, these two approaches are the only ones that used the common SNPs individually. Clearly, several SNPs in these sequence data, including those in our noncausal genes, have population-specific allele frequencies. Given that the genotype data were not simulated, we hypothesize that the inflated rates could be explained by the observed

Table 2 Type I error rates at α = 5% by gene in the unrelated individuals data set

Gene	SM[a]	T_{maf} = 0.01		T_{maf} = 0.05		WS	VT	CMC1	CMC2	CMC3
		CA	CP	CA	CP					
Unadjusted Q1										
CTSS	0.020	0.005	0.005	0.030	0.040	0.055	0.020	0.020	0.020	0.030
FAM73A	0.020	0.035	0.035	n/a	n/a	0.075	0.055	n/a	n/a	n/a
FOXI1	0.150	0.040	0.030	0.040	0.030	0.000	0.000	n/a	n/a	0.110
PTGFR	0.040	0.020	0.025	0.025	0.025	0.010	0.080	0.035	0.040	n/a
IFI44	0.350	0.055	0.050	0.110	0.140	0.040	0.120	0.305	0.305	0.220
MAPT	0.175	0.100	0.200	0.610	0.350	0.555	0.390	0.130	0.110	0.115
LY75	0.075	0.010	0.005	0.015	0.030	0.020	0.010	0.065	0.075	0.155
Q1 adjusted for the top five principal components										
CTSS	0.015	0.040	0.040	0.040	0.040	0.125	0.060	0.025	0.030	0.045
FAM73A	0.025	0.005	0.005	n/a	n/a	0.020	0.020	n/a	n/a	n/a
FOXI1	0.040	0.020	0.030	0.020	0.030	0.000	0.000	n/a	n/a	0.035
PTGFR	0.015	0.065	0.025	0.010	0.015	0.125	0.060	0.035	0.030	n/a
IFI44	0.075	0.030	0.025	0.010	0.015	0.010	0.015	0.020	0.020	0.000
MAPT	0.055	0.010	0.015	0.025	0.040	0.010	0.005	0.010	0.050	0.215
LY75	0.055	0.005	0.010	0.010	0.010	0.060	0.030	0.015	0.025	0.015

Estimates outside the 95% confidence interval are underlined. n/a, not applicable.

[a] Bonferroni-corrected *P*-value.

differences in the mean of Q1 between the four populations (−0.059, −0.002, 0.021, and 0.072 in Africans, Chinese, Japanese, and Europeans, respectively.

We recomputed the type I error accounting for possible clusters. First, we ran a principal components (PC) analysis with Eigenstrat [11] using the full mini-exome SNP data excluding SNPs with MAF < 5%. In each replicate, we computed the residual of Q1 obtained by regression of Q1 on the first five PCs. We reestimated the type I error levels using the residual of Q1 as the phenotype. The last 10 columns of Table 2 show the results. As can be seen, after adjusting for the five PCs, only a few of the type I error estimates remained higher than expected. In fact, most of the estimates were lower than expected.

In conclusion, to estimate the power of these approaches in the data sets, we used two strategies (Table 3): Power was first computed at a theoretical level of 5%, although the different approaches may not have comparable true false-positive rates; second, power was computed accounting for the five PCs, that is, using the residuals of Q1. All methods performed well for the *KDR* and *FLT1* genes. Conversely, all but two methods performed poorly (power < 10%) for two genes: For *ELAVL4* the power was greater than 30% using the SM and CMC3 approaches, and for *HIF3A* the power was greater than 17% for the CMC2 and CMC3 approaches. For the remaining four genes, one of the pooling methods outperformed the SM method after a Bonferroni correction. In these data, the CA and CP approaches had roughly similar power, and so, in what follows, the CP method will serve as a reference.

The choice of the threshold T_{maf} seems to have a large effect on power, and, in general, the power is higher when the criteria are less stringent (T_{maf} = 5% vs. 1%). Although this is not surprising for genes with causal SNPs having 1% < MAF < 5% (*ARNT*, *HIF1A*), we made the same observation for genes with all causal SNPs having a MAF < 1% (*FLT4* and *VEGFA*; see Table 1). This may suggest that allele correlation within these genes exists among causal and noncausal rare variants. The VT approach, which does not require a predefined choice on T_{maf}, did not appear to outperform the CP approach. On the other hand, one of the univariate (WS) or multivariate (CMC3) collapsing methods that uses all SNPs showed better power than the CP method. This again may be explained by allele correlation among SNPs. When adjusting for population stratification, again, all approaches had the greatest power for the *FLT1* and *KDR* genes and the lowest power for the *ELAVL4* and *HIF3A* genes. Nonetheless, most power estimates were lower, and the power drop was noticeable, especially for the *FLT4* and *HIF1A* genes. However, it is unclear whether this drop is fully explained by the lower values of the adjusted false-positive rates.

Table 3 Power rates at α = 5% by gene in the unrelated individuals data set

Gene	SM[a]	T_{maf} = 0.01		T_{maf} = 0.05		WS	VT	CMC1	CMC2	CMC3
		CA	CP	CA	CP					
Unadjusted Q1										
ARNT	0.86	0.04	0.04	0.79	0.83	0.53	0.76	0.93	0.96	0.94
ELAVL4	0.31	0.05	0.05	0.05	0.05	0.00	0.06	0.07	0.07	0.41
FLT1	0.99	0.85	0.91	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FLT4	0.33	0.41	0.38	0.65	0.62	0.78	0.76	0.50	0.47	n/a
HIF1A	0.42	0.07	0.07	0.62	0.59	0.45	0.51	0.62	0.62	n/a
HIF3A	0.02	0.03	0.02	0.07	0.07	0.06	0.04	0.20	0.17	0.10
KDR	0.96	0.97	0.99	1.00	1.00	1.00	1.00	0.99	0.99	1.00
VEGFA	0.26	0.13	0.13	0.41	0.44	0.54	0.45	0.31	0.31	n/a
VEGFC	0.58	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Q1 adjusted for the top five principal components										
ARNT	0.44	0.05	0.05	0.49	0.05	0.37	0.44	0.56	0.67	0.60
ELAVL4	0.07	0.07	0.07	0.07	0.07	0.05	0.12	0.06	0.06	0.01
FLT1	1.00	0.67	0.80	0.98	1.00	1.00	1.00	0.99	1.00	1.00
FLT4	0.09	0.03	0.02	0.04	0.03	0.01	0.02	0.04	0.06	n/a
HIF1A	0.13	0.08	0.08	0.00	0.01	0.01	0.01	0.19	0.19	n/a
HIF3A	0.03	0.05	0.03	0.01	0.00	0.00	0.00	0.03	0.04	0.03
KDR	0.74	0.63	0.74	0.84	0.85	0.99	0.93	0.72	0.69	0.78
VEGFA	0.25	0.13	0.13	0.04	0.06	0.19	0.32	0.08	0.10	n/a
VEGFC	0.56	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

Estimates outside the 95% confidence interval are underlined. n/a, not applicable.

[a] Bonferroni-corrected *P*-value.

Estimates of type I error and power rates in the family data set

Table 4 shows the type I error and power rates estimated at the gene level of each association approach for the family data set. It also shows the number of SNPs, causal and noncausal, that are polymorphic in the family samples. Type I error rates appeared to be better controlled in the family data than in the unrelated individuals data set with a few exceptions, especially the *MAPT* gene, for which most type I errors were biased upward. This gene is located in a genomic region with a low recombination rate and a long range of linkage disequilibrium. All association approaches show high and similar power rates for *VEGFA*. High power (>80%) was observed for *FLT1* using the SM and CP approaches and for *KDR* using the CA(0–5%), CP (0–5%), VT, and CMC1 approaches. In general, as observed in the unrelated individuals data set, the CA and CP approaches showed greater power under the less stringent T_{maf} criterion of 5% versus when T_{maf} = 1%.

Power of collapsing approaches in unrelated individuals versus family data set

Two causal genes (*FLT1*, *KDR*) were consistently detected with good power (>80%) in the unrelated individual and family data sets, irrespective of the association approach. One gene (*VEGFA*) was detected in the family sample but not in the unrelated individuals data set (power < 54%, or power < 32% after adjusting for population stratification). Conversely, *ARNT* was detected in the unrelated individuals data set (power = 96%, or power = 77% after adjusting for population stratification) but not in the family data (power = 12%).

Conclusions

We found that for some genes collapsing approaches may be powerful tools to detect multiple rare variants for complex traits. In particular, the choice of the threshold T_{maf} seems to have a large effect on power, and, in general, we found a higher power when the criterion was less stringent (T_{maf} = 5% vs. 1%). In the same vein, including all SNPs, whether by means of a univariate or a multivariate collapsing approach, can improve the power. In addition, a few of the causal genes were detected in both the related and the unrelated individuals data, whereas other causal genes were detected only in either the unrelated individuals or the family data. However, in these data the power of association was often limited. More important, we found that type I error rates may be highly variable between genes and between approaches.

We faced the challenge of assessing the proper type I error to validate our power comparisons. We acknowledge that our type I and type II error rates may not be generalized because of the way the GAW17 data were simulated: Phenotype but not genotype data were generated. Further, because the genotypes of founders did not vary between replicates, each family was either always

Table 4 Type I error and power at α = 5% by gene in family data set

Gene	*N*	*N* (*V*) with MAF < 5%	*N* (*V*) with MAF < 1%	SM[a]	T_{maf} = 0.01		T_{maf} = 0.05		WS	CMC1
					CA	CP	CA	CP		
Noncausal genes: type I error										
PTGFR	7	4 (0)	7 (0)	0.030	0.095	0.065	0.015	0.010	0.070	0.030
IFI44	9	7 (0)	8 (0)	0.060	0.030	0.025	0.030	0.040	0.010	0.175
FAM73A	3	3 (0)	3 (0)	0.025	0.020	0.020	0.015	0.020	0.035	n/a
MAPT	19	8 (0)	14 (0)	0.210	0.145	0.180	0.035	0.010	0.155	0.015
CTSS	3	2 (0)	2 (0)	0.020	0.015	0.015	0.015	0.015	0.020	n/a
FOXI1	5	3 (0)	3 (0)	0.020	0.020	0.055	0.055	0.055	0.045	0.000
LY75	49	30 (0)	39 (0)	0.055	0.070	0.045	0.030	0.035	0.120	0.035
Causal genes: power										
ARNT	7	6 (2)	4 (1)	0.04	0.04	0.03	0.01	0.01	0.12	0.03
ELAVL4	8	6 (1)	5 (1)	0.13	0.07	0.07	0.10	0.10	0.04	0.07
FLT1	16	13 (4)	8 (2)	0.95	0.02	0.02	0.57	0.82	0.44	0.33
FLT4	3	3 (0)	2 (0)	0.04	0.16	0.16	0.17	0.17	0.12	0.10
HIF1A	1	1 (1)	0 (0)	0.01	n/a	n/a	0.05	n/a	0.05	n/a
HIF3A	12	8 (1)	6 (1)	0.10	0.01	0.01	0.04	0.05	0.13	0.03
KDR	5	4 (4)	3 (3)	0.61	0.51	0.51	0.89	0.89	0.91	0.82
VEGFA	4	4 (1)	3 (1)	1	1	1	1	1	0.82	1
VEGFC	1	1 (1)	1 (1)	1	n/a	n/a	n/a	n/a	n/a	n/a

N, number of polymorphic SNPs. *V*, number of polymorphic causal variants.

[a] Bonferroni-corrected *P*-value.

informative (at least one founder carries a causal variant) or never informative (no founder carries a causal variant) for testing association to a given causal variant.

Finally, our results also raise an interesting point that might deserve future investigation, namely, that the collapsing and the single-locus association approaches may not be affected to the same extent by population stratification. Our results suggest that collapsing approaches may be more robust, especially in the presence of multiple variants.

Acknowledgments
The authors thank the French National Agency of Research (ANR-08-MNP-012). NB was funded by the European Community's Seventh Framework Programme ([FP7/2007- 2013] under grant agreement n° 212877 (UEPHA*MS)).
This article has been published as part of *BMC Proceedings* Volume 5 Supplement 9, 2011: Genetic Analysis Workshop 17. The full contents of the supplement are available online at http://www.biomedcentral.com/1753-6561/5?issue=S9.

Author details
[1]INSERM UMR1043, CPTP, CHU Purpan, Toulouse, 31024, France. [2]Université Paul Sabatier, Toulouse, France.

Authors' contributions
MS, ASP and MMacé performed the statistical analyses. MS, NB, and MMartinez drafted the manuscript. MMartinez conceived the study design and coordinated the study. All authors read and approved the final manuscript.

Competing interests
The authors declare that there are no competing interests.

Published: 29 November 2011

References

1. Altshuler D, Daly MJ, Lander ES: **Genetic mapping in human disease.** *Science* 2008, **322**:881-888.
2. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR: **Pooled association tests for rare variants in exon-resequencing studies.** *Am J Hum Genet* 2010, **86**:832-838.
3. Li B, Leal SM: **Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data.** *Am J Hum Genet* 2008, **83**:311-321.
4. Madsen BE, Browning SR: **A groupwise association test for rare mutations using a weighted sum statistic.** *PLoS Genet* 2009, **5**:e1000384.
5. Morris AP, Zeggini E: **An evaluation of statistical approaches to rare variant analysis in genetic association studies.** *Genet Epidemiol* 2009, **34**:188-193.
6. Zhu X, Feng T, Li Y, Lu Q, Elston RC: **Detecting rare variants for complex traits using family and unrelated data.** *Genet Epidemiol* 2010, **34**:171-187.
7. Almasy LA, Dyer TD, Peralta JM, Kent JW Jr, Charlesworth JC, Curran JE, Blangero J: **Genetic Analysis Workshop 17 mini-exome simulation.** *BMC Proc* 2011, **5(suppl 9)**:S2.
8. Dering C, Pugh E, Ziegler A: **Statistical analysis of rare sequence variants: an overview of collapsing methods.** *Genet Epidemiol* 2011, **X(suppl X)**:X-X.
9. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, *et al*: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81**:559-575.
10. Boerwinkle E, Chakraborty R, Sing C: **The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and analytical methods.** *Ann Hum Genet* 1986, **50**:181-194.
11. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies.** *Nat Genet* 2006, **38**:904-909.

doi:10.1186/1753-6561-5-S9-S33
Cite this article as: Saad *et al.*: Comparative study of statistical methods for detecting association with rare variants in exome-resequencing data. *BMC Proceedings* 2011 **5**(Suppl 9):S33.

OPEN ACCESS Freely available online

PLoS GENETICS

Genome-Wide Scan Identifies *TNIP1*, *PSORS1C1*, and *RHOB* As Novel Risk Loci for Systemic Sclerosis

Yannick Allanore[1,2]*, Mohamad Saad[3], Philippe Dieudé[4], Jérôme Avouac[1,2], Jorg H. W. Distler[5], Philippe Amouyel[6], Marco Matucci-Cerinic[7], Gabriella Riemekasten[8], Paolo Airo[9], Inga Melchers[10], Eric Hachulla[11], Daniele Cusi[12,13], H.-Erich Wichmann[14,15], Julien Wipff[1,2], Jean-Charles Lambert[6], Nicolas Hunzelmann[16], Kiet Tiev[17], Paola Caramaschi[18], Elisabeth Diot[19], Otylia Kowal-Bielecka[20], Gabriele Valentini[21], Luc Mouthon[22], László Czirják[23], Nemanja Damjanov[24], Erika Salvi[12,13], Costanza Conti[25], Martina Müller[26,27], Ulf Müller-Ladner[28], Valeria Riccieri[29], Barbara Ruiz[2], Jean-Luc Cracowski[30], Luc Letenneur[31,32], Anne Marie Dupuy[33], Oliver Meyer[4], André Kahan[1], Arnold Munnich[2], Catherine Boileau[2,34], Maria Martinez[3]

1 Université Paris Descartes, Rhumatologie A, INSERM, U1016, Hôpital Cochin, APHP, Paris, France, **2** INSERM, U781, Université Paris Descartes, Hôpital Necker, Paris, France, **3** INSERM, U563, CHU Purpan, Université Paul Sabatier, Toulouse, France, **4** Université Paris Diderot, Rhumatologie, INSERM, U699, Hôpital Bichat Claude-Bernard, Paris, France, **5** Department for Internal Medicine 3 and Institute for Clinical Immunology Friedrich-Alexander-University Erlangen-Nuremberg, Nuremberg, Germany, **6** INSERM, U744, Institut Pasteur de Lille, Université de Lille Nord, Lille, France, **7** Department of Biomedicine, Division of Rheumatology AOUC, Denothe Centre, University of Florence, Florence, Italy, **8** Department of Rheumatology and Clinical Immunology, Charité University Hospital, Berlin, Germany, **9** Rheumatology and Clinical Immunology, Spedali Civili, Brescia, Italy, **10** Clinical Research Unit for Rheumatology, University Medical Center, Freiburg, Germany, **11** Lille II University, Internal Medecine Department, Lille, France, **12** University of Milano, Department of Medicine, Surgery, and Dentistry San Paolo School of Medicine, **13** Genomics and Bioinformatics Platform, Fondazione Filarete, Milan, Italy, **14** Institute of Epidemiology I, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany, **15** Institute of Medical Informatics, Biometry, and Epidemiology, Chair of Epidemiology, Ludwig-Maximilians-Universität and Klinikum Grosshadern, Munich, Germany, **16** Department of Dermatology, University of Cologne, Köln, Germany, **17** Université Pierre et Marie Curie, Service de Médecine Interne, Hôpital Saint Antoine, Paris, France, **18** Rheumatology Unit, University of Verona, Verona, Italy, **19** INSERM, U618, IFR 135, CHU Bretonneau, Tours, France, **20** Department of Rheumatology and Internal Medicine, Medical University of Bialystok, Bialystok, Poland, **21** Department of Clinical and Experimental Medicine, Rheumatology Unit, Second University of Naples, Naples Italy, **22** Université Paris Descartes, Médecine Interne, Hôpital Cochin, APHP, Paris, France, **23** Department of Immunology and Rheumatology, University of Pécs, Pécs, Hungary, **24** Institute of Rheumatology, School of Medicine, University of Belgrade, Belgrade, Serbia, **25** Kos Genetic SRL, Milano, Italy, **26** Institute of Genetic Epidemiology, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany, and, **27** Institute of Medical Informatics, Biometry and Epidemiology, Chair of Epidemiology and Chair of Genetic Epidemiology, Ludwig-Maximilians-Universität and Department of Medicine I, University Hospital Grosshadern, Ludwig-Maximilians-Universität, Munich Germany, **28** University of Giessen, Department of Rheumatology and Clinical Immunology Kerckhoff-Klinik, Bad Nauheim, Germany, **29** Division of Rheumatology, Department of Internal Medicine and Medical Specialities, University "Sapienza," Rome, Italy, **30** INSERM, CIC3, CHU Grenoble, France, **31** INSERM, U897, Bordeaux, France, **32** Université Bordeaux Segalen, Bordeaux, France, **33** INSERM, U888, Hôpital de la Colombière, Montpellier, France, **34** Université Versailles-SQY, Laboratoire de Biochimie, d'Hormonologie et de Génétique Moléculaire, Hôpital Ambroise Paré, AP-HP, Boulogne, France

Abstract

Systemic sclerosis (SSc) is an orphan, complex, inflammatory disease affecting the immune system and connective tissue. SSc stands out as a severely incapacitating and life-threatening inflammatory rheumatic disease, with a largely unknown pathogenesis. We have designed a two-stage genome-wide association study of SSc using case-control samples from France, Italy, Germany, and Northern Europe. The initial genome-wide scan was conducted in a French post quality-control sample of 564 cases and 1,776 controls, using almost 500 K SNPs. Two SNPs from the MHC region, together with the 6 loci outside MHC having at least one SNP with a $P<10^{-5}$ were selected for follow-up analysis. These markers were genotyped in a post-QC replication sample of 1,682 SSc cases and 3,926 controls. The three top SNPs are in strong linkage disequilibrium and located on 6p21, in the *HLA-DQB1* gene: rs9275224, $P=9.18\times10^{-8}$, OR = 0.69, 95% CI [0.60–0.79]; rs6457617, $P=1.14\times10^{-7}$ and rs9275245, $P=1.39\times10^{-7}$. Within the MHC region, the next most associated SNP (rs3130573, $P=1.86\times10^{-5}$, OR = 1.36 [1.18–1.56]) is located in the *PSORS1C1* gene. Outside the MHC region, our GWAS analysis revealed 7 top SNPs ($P<10^{-5}$) that spanned 6 independent genomic regions. Follow-up of the 17 top SNPs in an independent sample of 1,682 SSc and 3,926 controls showed associations at *PSORS1C1* (overall $P=5.70\times10^{-10}$, OR:1.25), *TNIP1* ($P=4.68\times10^{-9}$, OR:1.31), and *RHOB* loci ($P=3.17\times10^{-6}$, OR:1.21). Because of its biological relevance, and previous reports of genetic association at this locus with connective tissue disorders, we investigated TNIP1 expression. A markedly reduced expression of the *TNIP1* gene and also its protein product was observed both in lesional skin tissue and cultured dermal fibroblasts from SSc patients. Furthermore, TNIP1 showed *in vitro* inhibitory effects on inflammatory cytokine-induced collagen production. The genetic signal of association with *TNIP1* variants, together with tissular and cellular investigations, suggests that this pathway has a critical role in regulating autoimmunity and SSc pathogenesis.

Citation: Allanore Y, Saad M, Dieudé P, Avouac J, Distler JHW, et al. (2011) Genome-Wide Scan Identifies *TNIP1*, *PSORS1C1*, and *RHOB* As Novel Risk Loci for Systemic Sclerosis. PLoS Genet 7(7): e1002091. doi:10.1371/journal.pgen.1002091

Editor: Mark I. McCarthy, University of Oxford, United Kingdom

Received October 28, 2010; **Accepted** April 5, 2011; **Published** July 7, 2011; 11 February 2011

Funding: This project was funded by Agence Nationale pour la Recherche (Project ANR-08-GENO-016-1) and supported by research grants from SERVIER research group, SANOFI-AVENTIS, Association des Sclérodermies de France, and Groupe Français de Recherche sur la Sclérodermie. The KORA (Cooperative health research in the Region of Augsburg) studies were financed by the Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany, and supported by grants from the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. Part of this work was financed by the German National Genome Research Network (NGFN). This research was supported within the Munich Center of Health Sciences (MC Health) as part of LMUinnovativ. HYPERGENES (European Network for Genetic-Epidemiological Studies) is funded by EU within the FP7: HEALTH-F4-2007-201550. The sample considered in the present study has been collected by the Milano group in collaboration with the Center of Transfusion Medicine, Cellular Therapy, and Cryobiology, Foundation Ca' Granda Ospedale Maggiore Policlinico, and Fondazione Filarete. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: yannick.allanore@cch.aphp.fr

Introduction

Systemic sclerosis (MIM181750) is a connective tissue disease characterized by generalized microangiopathy, severe immunologic alterations and massive deposits of matrix components in the connective tissue. Being an orphan disease, SSc presents a major medical challenge and is recognized as the most severe connective tissue disorder with high risk of premature deaths [1]. Epidemiological data on SSc vary in different parts of the world and depend on selection criteria for the study population. Inasmuch, the prevalence of the disease fluctuates across global regions and population-based studies result in higher prevalence than do hospital records-based studies. In North America, the prevalence of SSc has been reported as 0.7–2.8 per 10,000 in a Canadian study, whereas in the U.S. figures of 2.6 per 10,000 versus 7.5 per 10,000 were reported by medical records - versus population-based studies, respectively. In Europe, a prevalence of 1.6 per 10,000 was reported in Denmark, 3.5 per 10,000 in Estonia, 1.58 per 10,000 adults (95% confidence interval, 129–187) in Seine-Saint-Denis in France [2–4]. The risk of SSc is increased among first-degree relatives of patients, compared to the general population. In a study of 703 families in the US, including 11 multiplex SSc families, the familial relative risk in first-degree relatives was about 13, with a 1.6% recurrence rate, compared to 0.026% in the general population [5]. The sibling risk ratio was about 15 (ranging from 10 to 27 across cohorts). The only twin study reported to date included 42 twin pairs [6]. The data showed a similar concordance rate in monozygotic twins (4.2%, n = 24) and dizygotic twins (5.6%) (NS) and an overall cross-sectional concordance rate of 4.7%. However, concordance for the presence of antinuclear antibodies was significantly higher in the monozygotic twins (90%) than in the dizygotic twins (40%) suggesting that genetics may be important for the auto-immune part of the disease.

The aetiology of SSc is still unclear but some key steps have been described, in particular early endothelial damage and dysregulation of the immune system with abnormal autoantibody production [7]. At the cellular level, early events include endothelial injury and perivascular inflammation with the release of a large array of inflammatory mediators [8,9]. In the advanced stage, a progressive activation of fibroblasts in the skin and in internal organs leads to hyperproduction of collagen and irreversible tissue fibrosis [9]. Epidemiological investigations indicate that SSc follows a pattern of multifactorial inheritance [10]. Previous candidate-gene association studies have only identified a handful of SSc risk loci, most contributing to the genetic susceptibility of other autoimmune diseases [9–16]. So far, two genome-wide association studies of SSc have been conducted [17,18]. The studies differ according to the ancestry of the studied population (Korean vs US/European) and the genome-wide association data: map density (~440 K vs 280 K SNPs) and sample size (~700 vs ~7300 subjects). They provided evidence of association with known MHC loci, but only one 'new' locus was identified at *CD247* in the US/European dataset, variants at *CD247* being known to contribute to the susceptibility of systemic lupus erythematosus [18].

The diagnosis of SSc is based on recognized clinical criteria established decades ago however, these do not include specific autoantibodies or recent tools for assessment of the disease [19,20]. Therefore, phenotypic heterogeneity is a concern for SSc and genetic heterogeneity is also highly probable with regards to data obtained in other connective tissue disorders. Given these considerations, and previous findings in other autoimmune diseases, it is apparent that additional risk variants for SSc remain to be discovered. Therefore, to identify further common variants that contribute to SSc risk in the European population, we conducted a two-stage GWAS, in two case-control samples (total >8,800 subjects).

Results/Discussion

We established a collaborative consortium including groups from 4 European countries (France, Italy, Germany and Eastern-Europe) from which we were able to draw upon a combined sample of over 8,800 subjects (before quality control) and conducted a two-stage genome-wide association study. In stage 1, we genotyped 1,185 samples on Illumina Human610-Quad BeadChip and genotypes obtained using the same chip from 2,003 control subjects were made available to us from the 3C study [21,22]. After stringent quality control, we finally tested for association in stage-1, 489,814 autosomal SNPs in 2,340 subjects (564 cases and 1,776 controls) (Table 1). We tested for association between each SNP and SSc using the logistic regression association test, assuming additive genetic effects. The quantile-quantile plot and estimation of the genomic inflation factor ($\lambda = 1.035$) indicated minimal overall inflation (Figure 1A). The genome-wide logistic association results are presented in Figure 1B. Table S1 provides details for all SNPs with $P<10^{-4}$, including one SNP exceeding $P<10^{-7}$, the Bonferroni threshold for genome-wide significance. The three top SNPs were located on 6p21, in the *HLA-DQB1* gene: rs9275224, $P = 9.18\times10^{-8}$, OR = 0.69, 95%CI[0.60–0.79]; rs6457617, $P = 1.14\times10^{-7}$ and rs9275245, $P = 1.39\times10^{-7}$ (Figure 1B and Table 2). Several associated SNPs in *HLA-DQB1* have already been reported but rs6457617 was also

Author Summary

Systemic sclerosis (SSc) is a connective tissue disease characterized by generalized microangiopathy, severe immunologic alterations, and massive deposits of matrix components in the connective tissue. Epidemiological investigations indicate that SSc follows a pattern of multifactorial inheritance; however, only a few loci have been replicated in multiple studies. We undertook a two-stage genome-wide association study of SSc involving over 8,800 individuals of European ancestry. Combined analyses showed independent association at the known *HLA-DQB1* region and revealed associations at *PSORS1C1*, *TNIP1*, and *RHOB* loci, in agreement with a strong immune genetic component. Because of its biological relevance, and previous reports of genetic association at this locus with other connective tissue disorders, we investigated TNIP1 expression. We observed a markedly reduced expression of the gene and its protein product in SSc, as well as its potential implication in control of extra-cellular matrix synthesis, providing a new clue for a link between inflammation/immunity and fibrosis.

identified as the most associated SNP in the previous US/European GWAS study [18]. Of note, the three SNPs in *HLA-DQB1* are in strong LD ($r^2>0.97$). Within the MHC region, the next most associated SNP (rs3130573, $P=1.86\times10^{-5}$, OR = 1.36[1.18–1.56]) is located in the psoriasis susceptibility 1 candidate 1 (*PSORS1C1*) gene (Table 2), a candidate gene for psoriasis [23]. Conditional analyses of susceptibility variants within MHC showed that there were two independent association signals at rs6457617 (*HLA-DQB1*) and at rs3130573 (*PSORS1C1*). Indeed, the association at PSORS1C1 remained significant ($P<2.1\times10^{-5}$) after controlling for the association at HLA-DQB1 and the association at HLA-DQB1 remained also significant ($P<1.5\times10^{-7}$) after controlling for the association at PSORS1C1 (Table S2).

Outside the MHC region, our GWAS analysis revealed 7 top SNPs ($P<10^{-5}$) that spanned 6 independent genomic regions (Figure 1B and Table S1). Conditional analyses of each of them on *HLA-DQB1* showed no significant drop in the association signals (Table S3). The 6 loci having at least one SNP with a $P<10^{-5}$ were selected for follow-up analysis. Within each locus we selected the SNPs with the strongest ($P<10^{-4}$) association signals to be genotyped in a post-QC replication sample of 1,682 SSc cases and 3,926 controls (Table 1). To this list we added two top SNPs in HLA-DQB1 and the SNP in PSORS1C1. Finally, we further included 4 SNPs at the two known loci (*STAT4* and *TNPO3-IRF5*) and at the newly identified locus (*CD247*) by Radstake et al [18]. Out of a total set of 21 SNPs submitted for replication, 20 passed the quality-control analyses.

Stratified association analyses in stage 2 data (Table 2), confirmed the strong association for *HLA-DQB1* (rs6457617, $P=1.35\times10^{-28}$) at 6p21.3 and also with the *PSORS1C1* variant (rs3130573, $P=4.98\times10^{-3}$) at 6p21.1. Of the 6 remaining loci selected in stage 1, only 2 were replicated with nominal P< 5% and with same direction of effect. They mapped at 2p24 (rs342070, P = 0.026; rs13021401, P = 0.024) and 5q33 (rs3792783, $P=4.14\times10^{-5}$; rs2233287, $P=4.38\times10^{-3}$; rs4958881, $P=2.09\times10^{-3}$). None of the replicated SNPs showed evidence for heterogeneity of effects among the 4 geographical origins (Breslow-day P>0.10). As expected, ORs estimated in the discovery tended to be higher than those obtained in the replication stage data. Afterwards, association signals from joint analyses of the 2 datasets (Table 2) consistently showed highly significant association for *HLA-DQB1* ($P=2.33\times10^{-37}$), *PSORS1C1* ($P=5.70\times10^{-10}$) and *TNIP1* ($P=4.68\times10^{-9}$), and also showed some evidence of association for *RHOB* ($P=3.17\times10^{-6}$). All populations showed same direction of effects (Figure 2). Finally, we also replicated association signals at *IRF5* ($P=3.49\times10^{-5}$; combined-$P=4.13\times10^{-7}$), at STAT4 ($P=1.9\times10^{-10}$; combined-$P=2.26\times10^{-13}$) and at the recently identified new SSc risk locus, *CD247* ($P=2.90\times10^{-5}$; combined-$P=1.30\times10^{-6}$) (Table 2). In our combined data, the locus-specific PAR estimates were 24% for *HLA-DQB1*,

Table 1. Description of the study population (post quality control samples).

	CASES					CONTROLS			TOTAL
	N	Mean age ±SD (years)	Female (%)	DcSSc	Topo+/ ACA+ (%)	N	Mean age ±SD (years)	Female (%)	
Stage 1: Discovery sample									
	564	56.6±17.4	84%	34.8%	26% / 38.3%	488 (Genesys)	49.2±11.7	80%	
						1288 (3C)	73.99±5.6	65.5%	
Total DIS						1776	69±11.9	69.9%	2340
Stage 2: Follow up samples									
French	370	58.8±14.5	82.9%	28.7%	27.8% / 50.5%	1906	38.6±21.2	55.0%	
Italian	596	56.3±13.4	88.6%	25.4%	32.7% / 46%	490 (Italian network)	47.9±13.2	84.6%	
						721 (Hypergenes)	59.1±6.7	43.3%	
Eastern	151	53.1±12.4	93.6%	45.7%	24.6% / 13.8%	148	30.5±11.6	50%	
German	565	56.6±13.9	88.5%	35%	32.4% / 38.4%	180 (German network)	55±16.3	50%	
						481 (KORA study)	63.1±7.25	48.8%	
Total REP	1682	56.81±20.9	86.5%	30.5%	31.2% / 42.8%	3926	50.6±23.2	52.5%	5608

DcSSc: diffuse cutaneous systemic sclerosis; TOPO: anti topoisomerase I antibodies; ACA: anti-centromere antibodies.
doi:10.1371/journal.pgen.1002091.t001

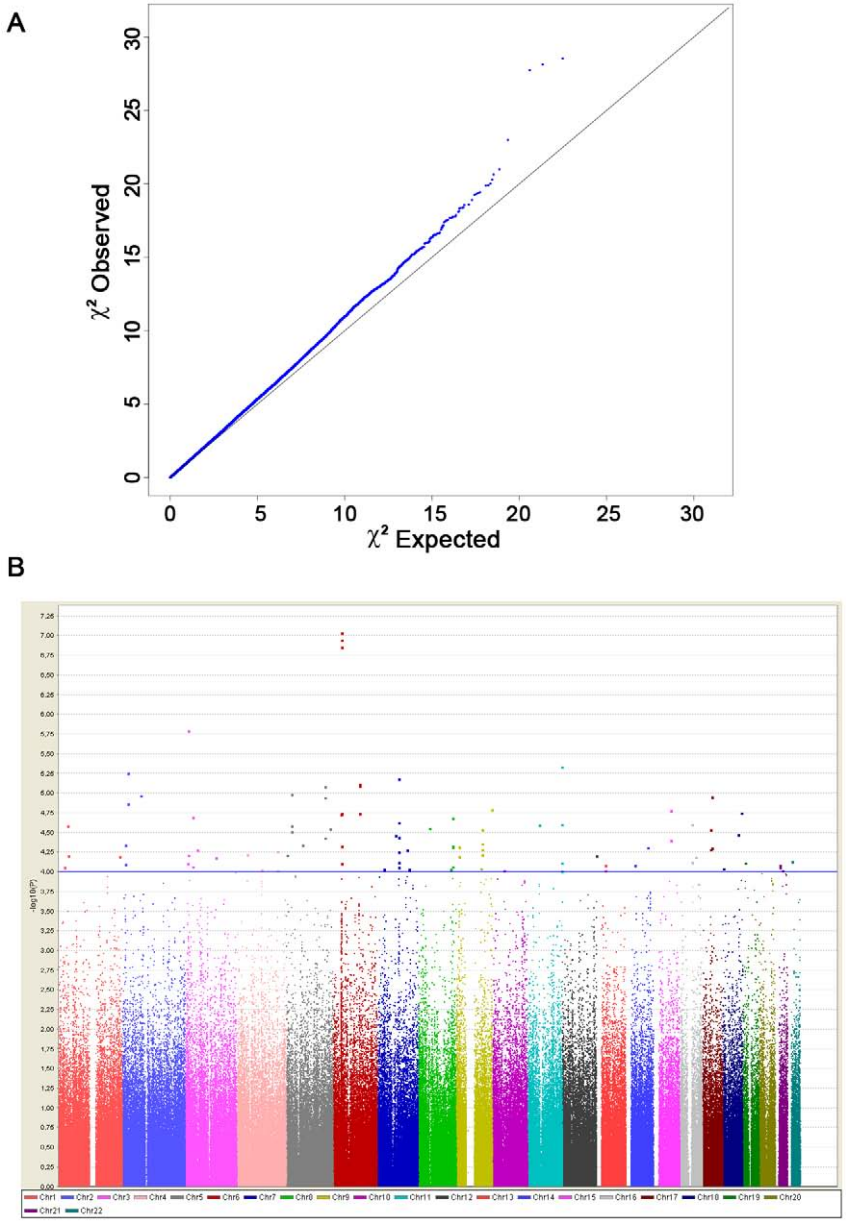


Figure 1. Genome-wide association results from the discovery phase. (A) Quantile-quantile plot for test statistics (logistic regression test) for 489,814 SNPs passing quality control. The plot shows a close match to the test statistics expected under the null distribution ($\lambda = 1.03$). (B) Manhattan plot representing the P values across the genome. The $-$log10 P of the logistic regression test (y axis) from 489,814 SNPs in 564 systemic sclerosis patients and 1,776 controls is plotted against its physical position (x-axis) on successive chromosomes. 90 SNPs with $P<10^{-4}$ lie above the blue horizontal line and are listed in Table S1. Highly significant association was observed with SNPs within the MHC locus, including 1 SNPs that reached the conservative threshold for genome-wide significance ($P<10^{-7}$).
doi:10.1371/journal.pgen.1002091.g001

4% for TNIP1, 8% for *PSORS1C1*, 7% for *CD247*, 8% for *STAT4* and 3% with *IFR5/TNPO3*. The combined PAR estimate was 47.4%.

As secondary analyses, we assessed homogeneity of SNP's effect between sub-categories of SSc (cutaneous sub-types and auto-antibodies). Case-only analyses revealed no significant evidence for heterogeneous ORs between cutaneous sub-types of SSc patients for any of the 5 replicated SNPs at 2p24 or 5q33 loci (Table S4A). Indeed, similar association signals were obtained from case-category association analyses (Table S4B). Altogether, the results did not suggest that the association signals in the newly identified 5q33 locus were driven by a specific sub-type of SSc. Conversely, for *HLA-DQB1* and *PSORS1C1* we found evidence of heterogeneity in OR estimates in positive *vs* negative ACA or TOPO auto-antibody SSc patients (Table S4A). Yet, the association signals in each of these sub-types of patients remained strong (Table S4B). These results support the previously reported hypothesis that the magnitude of the *HLA-DQB1* effect on SSc susceptibility may depend on auto-antibody status [11]. The GWAS stage had 78% power to detect loci of the effect sizes observed in the discovery sample for *TNIP1* variants (OR = 1.50) at a significance of $P<10^{-5}$. However, it is widely acknowledged that effect sizes of

Table 2. Genome-wide association and replication for systemic sclerosis risk variants.

Chr. (closest gene)	Pos. (bp)	SNP	Minor/ Major	MAF Cases/ Controls	P	OR	95%CI	MAF Cases/ Controls	$P	OR	95%CI	**P	$P	OR	95%CI
MHC loci															
6p21(PSORS1C1)	31 214 247	rs3130573	G/A	0.391/ 0.321	1.86E-05	1.36	(1.18–1.56)	0.416/0.373	4.98E-03	1.13	(1.04–1.23)	4.8E-01	5.70E-10	1.25	(1.17–1.35)
6p21 (HLA-DQB1)	32 767 856	rs9275224	A/G	0.405/ 0.496	9.18E-08	0.69	(0.6–0.79)	&NA	-	-	-	-	-	-	-
6p21 (HLA-DQB1)	32 771 829	rs6457617	C/T	0.408/ 0.498	1.14E-07	0.69	(0.6–0.79)	0.345/0.463	1.35E-28	0.61	(0.56–0.67)	1.0E-01	2.33E-37	0.62	(0.58–0.67)
Non MHC loci															
2p24 (RHOB)	20 548 952	rs342070	C/T	0.293/ 0.226	5.56E-06	1.42	(1.22–1.65)	0.258/0.235	2.61E-02	1.12	(1.01–1.23)	1.9E-01	4.66E-06	1.20	(1.11–1.30)
	20 552 000	rs13021401	T/C	0.289/ 0.225	1.37E-05	1.40	(1.2–1.63)	0.257/0.232	2.47E-02	1.12	(1.01–1.24)	1.3E-01	3.17E-06	1.21	(1.12–1.31)
3p25 (PPARG/TSEN2)	12 468 347	rs9855622	T/C	0.145/ 0.096	1.64E-06	1.66	(1.35–2.05)	0.097/0.109	9.86E-01	1.00	(0.85–1.17)	7.6E-03	1.05E-01	1.11	(0.94–1.17)
	12 234 616	rs310746	C/T	0.121/ 0.08	6.15E-05	1.55	(1.25–1.91)	0.074/0.077	8.69E-02	0.88	(0.77–1.02)	9.6E-01	4.22E-01	1.05	(0.98–1.25)
5q33 (TNIP1)	150 430 429	rs4958881	C/T	0.166/ 0.115	8.26E-06	1.54	(1.28–1.87)	0.151/0.130	4.38E-03	1.21	(1.06–1.38)	3.2E-01	5.79E-06	1.29	(1.17–1.42)
	150 435 925	rs3792783	G/A	0.208/ 0.152	1.14E-05	1.47	(1.24–1.75)	0.198/0.166	2.09E-03	1.21	(1.07–1.36)	6.8E-01	5.73E-07	1.29	(1.20–1.43)
	150 420 290	rs2233287	A/G	0.139/ 0.096	3.71E-05	1.55	(1.26–1.91)	0.121/0.103	4.14E-05	1.26	(1.13–1.40)	6.8E-01	4.68E-09	1.31	(1.15–1.43)
6p16-q16 (ASCC3)	101 444 332	rs9498419	A/G	0.522/ 0.446	7.71E–06	1.37	(1.19–1.57)	0.458/0.475	1.18E-01	0.93	(0.85–1.02)	4.6E-01	3.15E-01	1.04	(0.97–1.11)
	101 445 699	rs6919745	T/C	0.522/ 0.447	8.14E-06	1.37	(1.19–1.57)	0.461/0.478	7.47E-02	0.93	(0.85–1.01)	4.0E-01	3.34E-01	1.04	(0.97–1.11)
7p12-q21 (SEMA3A/HMG17P1)	84 166 013	rs4329228	C/A	0.305/ 0.239	6.66E-06	1.42	(1.22–1.65)	0.249/0.248	5.08E-01	0.97	(0.87–1.07)	3.4E-01	2.05E-01	1.06	(1.04–1.20)
	83 976 940	rs1029541	T/C	0.288/ 0.227	2.37E-05	1.39	(1.2–1.63)	0.223/0.231	7.93E-01	1.01	(0.92–1.12)	9.6E-01	1.50E-02	1.11	(0.97–1.15)
11q25 (OPCML)	132 284 603	rs2725466	G/A	0.403/ 0.328	4.60E-06	1.39	(1.21–1.59)	0.378/0.375	5.71E-01	0.98	(0.89–1.06)	6.2E-01	3.55E-03	1.11	(1.04–1.20)
	132 287 033	rs2725437	C/T	0.404/ 0.335	2.52E-05	1.35	(1.17–1.54)	0.395/0.387	8.33E-01	0.99	(0.91–1.08)	5.4E-01	1.82E-03	1.12	(1.04–1.20)
	132 300 779	rs10894623	T/G	0.317/ 0.256	7.75E-05	1.34	(1.16–1.55)	0.275/0.277	7.67E-01	0.99	(0.90–1.08)	1.4E-01	5.83E-02	1.08	(1.00–1.16)
Previously reported with $P<10^{-7}$															
1q22-23 (CD247)	165 687 049	rs2056626	G/T	0.354/0.393	1.70E-02	0.84	(0.73–0.97)	0.36/ 0.4	2.90E-05	0.82	(0.75–0.9)	3.8E-01	1.30E-06	0.83	(0.77–0.89)
2q32 (STAT4)	191 611 003	rs3821236	A/G	0.227/0.203	8.70E-02	1.16	(0.98–1.36)	0.24/ 0.2	2.10E-07	1.33	(1.2–1.49)	8.0E-01	2.09E-07	1.27	(1.16–1.39)
	191 672 878	rs7574865	T/G	0.272/0.219	2.50E-04	1.33	(1.14–1.55)	0.29/ 0.22	1.90E-10	1.40	(1.26–1.56)	9.0E-01	2.26E-13	1.38	(1.27–1.5)
7q32 (TNPO3-IRF5)	128 381 419	rs10488631	C/T	0.124/0.093	2.50E–03	1.39	(1.12–1.72)	0.14/ 0.10	3.49E-05	1.34	(1.17–1.54)	4.5E-01	4.13E-07	1.35	(1.2–1.51)

doi:10.1371/journal.pgen.1002091.t002

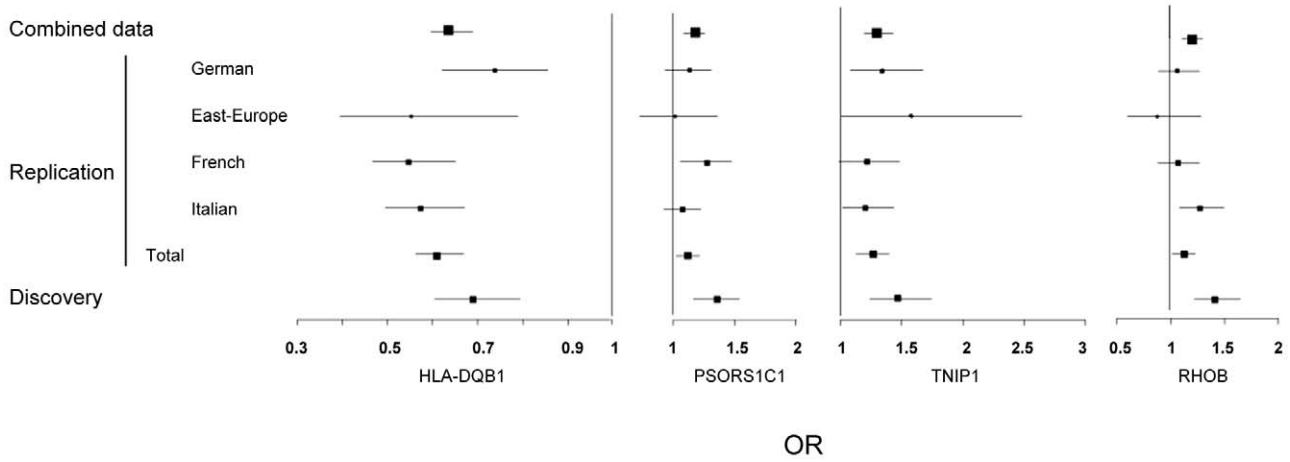


Figure 2. Forest plots showing odds ratios and confidence intervals of the *HLA-DQB1*, *PSORS1C1*, *TNIP1*, and *RHOB* associations in the various populations studied in stage-1 and stage-2 data.
doi:10.1371/journal.pgen.1002091.g002

significant GWAS loci are overestimates of true effects and other genes of lower effect sizes are unlikely to reach stringent significant thresholds.

Our GWAS analysis revealed strong association with *PSORS1C1*, which is ~1 Mb of *HLA-DQB1*. Notably, *PSORS1C1* is known to be involved in autoimmune response [23]. In the combined data, association with *PSORS1C1* was highly significant ($P = 5.70 \times 10^{-10}$) and remained significant after controlling for the association at HLA-DQB1. Altogether, our results suggest that this region is likely to contain more than one gene playing a role in the pathogenesis of autoimmune disorders [23,24]. Fine mapping at this locus is warranted to identify causal variants.

The three strongly associated SNPs at the 5q33 locus are located within the *TNFAIP3* interacting protein 1 (*TNIP1*) gene. *TNIP1* is a very interesting new candidate gene for SSc. The protein encoded by this gene exerts a negative regulation of NF-kappaB via two sequential activities: deubiquitination of Lys63-based chains and synthesis of Lys48-based chains on the TNF receptor-interacting protein and also inhibition of NF-KappaB processing [25]. TNIP1 interacts with A20 (TNFAIP3) to negatively regulate NF-kappaB. Several recent studies have suggested that the activation of some inflammatory factors may upregulate fibrotic mediators through Toll-like receptors (TLRs), thereby contributing to SSc pathogenesis [8]. It has been shown that TLR engagement leads to A20 induction in macrophages and that TNIP1/A20 is essential for the termination of TLR-induced NF-kappaB activity and proinflammatory cytokine production [26]. Although interactions between TNIP1 and A20 are not well known, A20 also acts as a deubiquitinating enzyme, suggesting a molecular link between deubiquitinating activity and the regulation of TLR signals [26]. Therefore, TNIP1 and A20 may play a critical role in the regulation of downstream TLR signals, and this issue will have to be addressed in SSc. Interestingly, variants at *TNIP1* have been shown to be implicated in systemic lupus erythematosus susceptibility [27,28] and in psoriasis [29]. Furthermore, we have recently reported an association of one *TNFAIP3* variant with SSc [30]. In our stage-1 data, evidence of association at *TNFAIP3* was nominal (lowest $P = 0.047$) and no pairwise interaction was found ($P > 0.06$) between *TNFAIP3* and *TNIP1* variants. Analysis of the LD structure across the *TNIP1* gene revealed that the 3 strongly associated SNPs belong to the same LD-block (Figure 3). No residual association signals were observed when rs3792783 and each of the other 2 SNPs were paired in conditional analyses. Therefore, any of them, or other variants yet to be identified, could be the causal variant(s). Interestingly, rs3792783 is located upstream from the transcription start site in exon 2 (Figure 3). It is noteworthy that previously reported lupus *TNIP1* variants were located in the same LD-block [27,28]. Because of the compelling evidence of the potential role of NF-kappaB in autoimmune diseases and our raised new signal association for SSc at *TNIP1* (a negative regulator of this pathway) we performed *ex vivo* and *in vitro* investigations to assess TNIP1 expression in SSc patients and healthy controls. For SSc patients, the results showed a strikingly reduced expression of TNIP1 in skin tissue (Figure 4A), and of both mRNA (Figure 4B) and protein (Figure 4C) synthesis by cultured dermal fibroblasts. Addressing the question of the potential link between the NF-kappaB pathway and the fibrotic propensity that characterizes SSc, we next assessed the influence of pro-inflammatory cytokines and TNIP1 on the synthesis of extra-cellular matrix by dermal fibroblasts in culture. Using cells from the skin of healthy controls (Figure 5) and SSc patients (Figure 6), we showed that recombinant TNIP1 abrogated collagen synthesis induced by inflammatory cytokines both at the mRNA and protein levels. It must be acknowledged that TNIP1 is described as an intra-cellular protein whereas we used recombinant protein added to cell supernatant in these experiments. The observed effects may be related to different hypotheses. TNIP1 has been described as a nuclear shuttling protein and it could have a chaperon-like activity, highly interacting with other protein that could result in engulfment of TNIP1 through interaction with a cell surface protein. Such intra-cellular effects of extra-cellular proteins has been shown also for the S100 family of proteins that have no leader sequence and for clusterin for which it is postulated that the protein could be taken up by interacting with either a yet unidentified receptor or by a mechanism related to their chaperon-like activity [31]. More work is needed to determine which of these hypotheses has to be retained and to investigate more in depth soluble TNIP1 In this first attempt to explore TNIP1 functional disturbances, we could not investigate a relationship between specific TNIP variants and *in vitro* or *in vivo* changes; this will need to be addressed ideally after the identification of the causal variant and using a much larger sample size. Nevertheless, our results raise a potential relationship between inflammation and fibrosis and open a new and highly

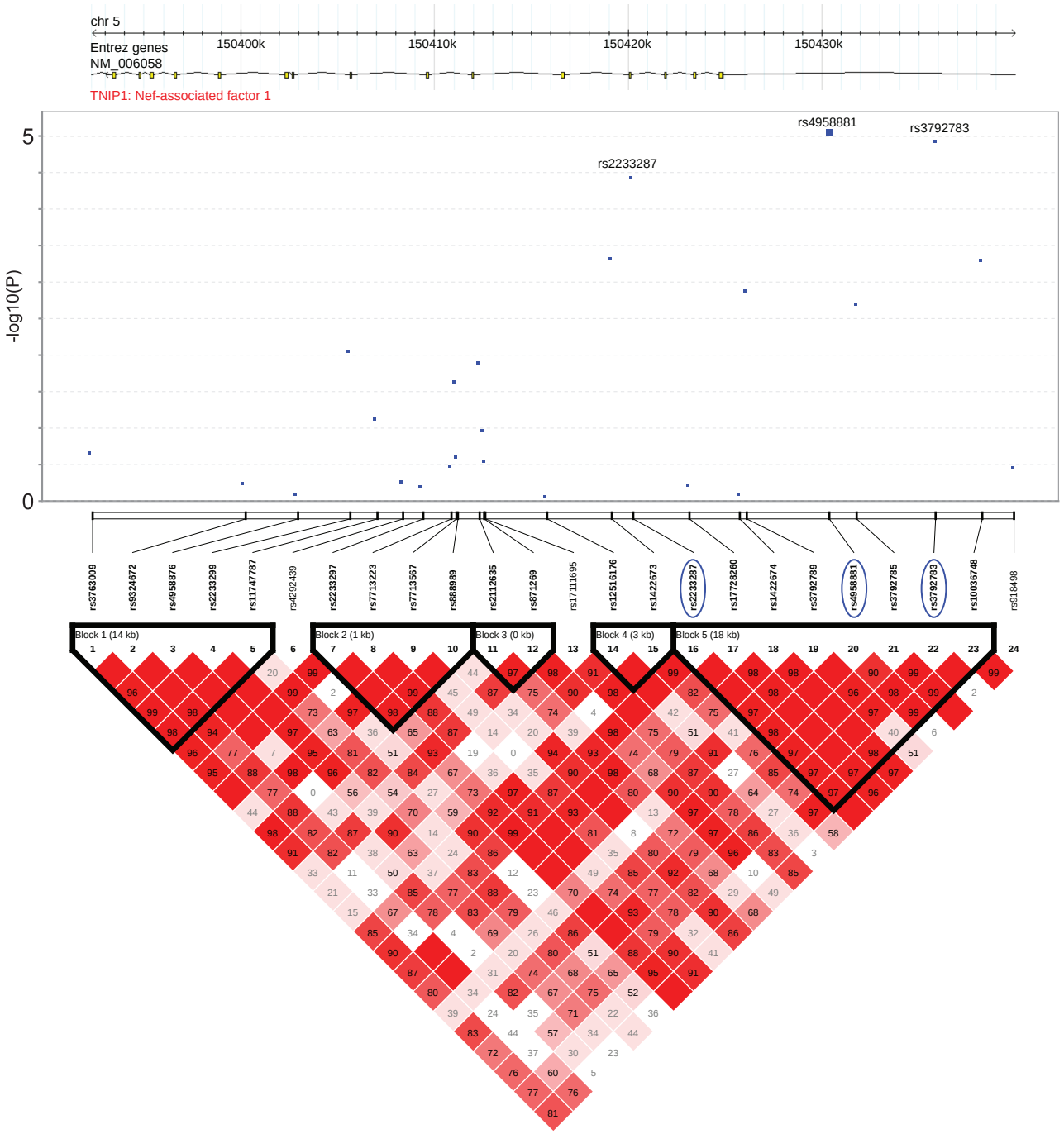


Figure 3. Association and linkage disequilibrium patterns at the TNIP1 gene. (A) Association of SNPs in TNIP1: −log10 P of the logistic regression test for association (y axis) in the GWAS stage of SNPs is plotted against their physical position. The continuous line corresponds to P<10−5, the minimum P value of the top 7 SNPs identified in stage 1. The three SNPs that were followed and replicated in stage 2 are highlighted by a blue circle. Positions are given as NCBI build. (B) Linkage disequilibrim patterns at the TNIP1gene: pairwise LD (D′) are indicated by color gradients: D′≥0.80, red; 0.5≤D′<0.8, pink; 0.2≤D′<0.5, light pink; D′<0.2, white. The 3 SNPs are in strong LD (r2 = 0.57/0.72 between rs3792783 and rs2233287/rs495881). Intron and exon structure of the *TNIP1* gene are taken from the UCSC Genome Browser.
doi:10.1371/journal.pgen.1002091.g003

relevant field of investigation in SSc pathogenesis and in fibrotic disorders.

Our next most associated SNPs at 2p24 are in strong LD ($r^2 = 0.98$) and map ~30 kb from *RHOB*. RHOB is the Ras homolog gene family member B that regulates protein signalling and intracellular protein trafficking. RhoB is essential for activity of farnesyltransferase inhibitors and also statins that are two strong potential future drugs in SSc [9,32]. To our knowledge, association to RHOB has never been reported so far. The signal for association was weaker at this locus and therefore will need to be confirmed in other samples and more investigations are warranted to assess RHOB implication in this disease.

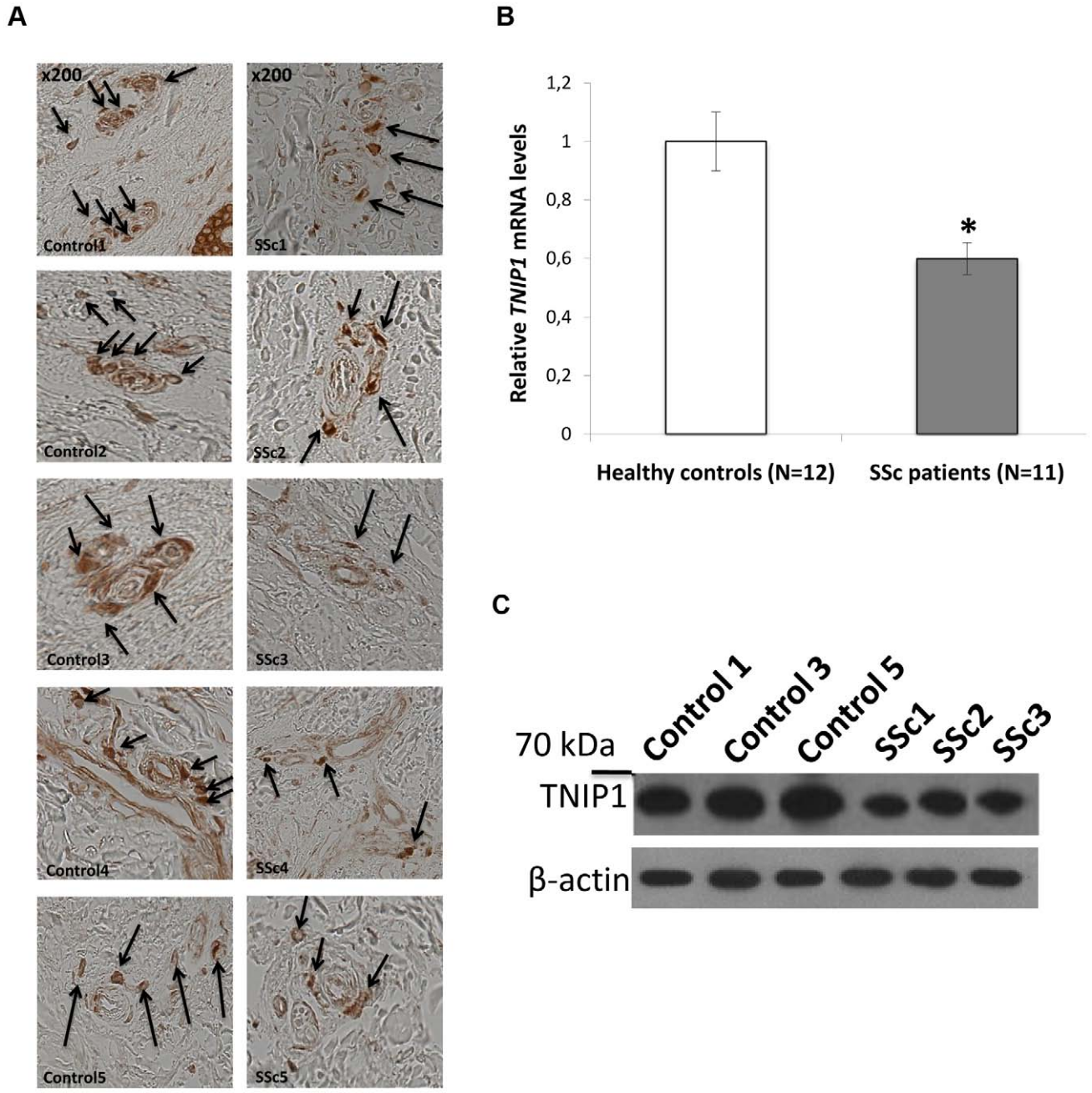


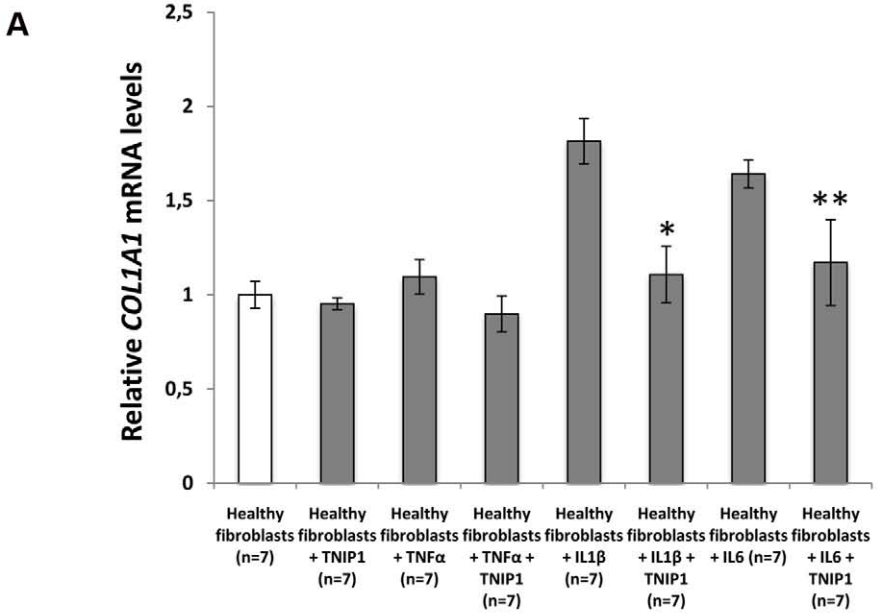
Figure 4. Decreased TNIP1 expression in SSc patients. (A) Expression of the TNIP1 protein was decreased ex vivo in SSc lesional skin tissue compared to controls (arrows indicate TNIP-1 positive cells). Shown are representative sections of the 5 patients and controls included in the analysis. (B) Consistent with these findings, a 1.7-fold decrease of *TNIP1* mRNA levels was observed in dermal fibroblasts from SSc patients (* indicates a *P*-value = 0.001 *vs.* controls). These results were confirmed at the protein level (C).
doi:10.1371/journal.pgen.1002091.g004

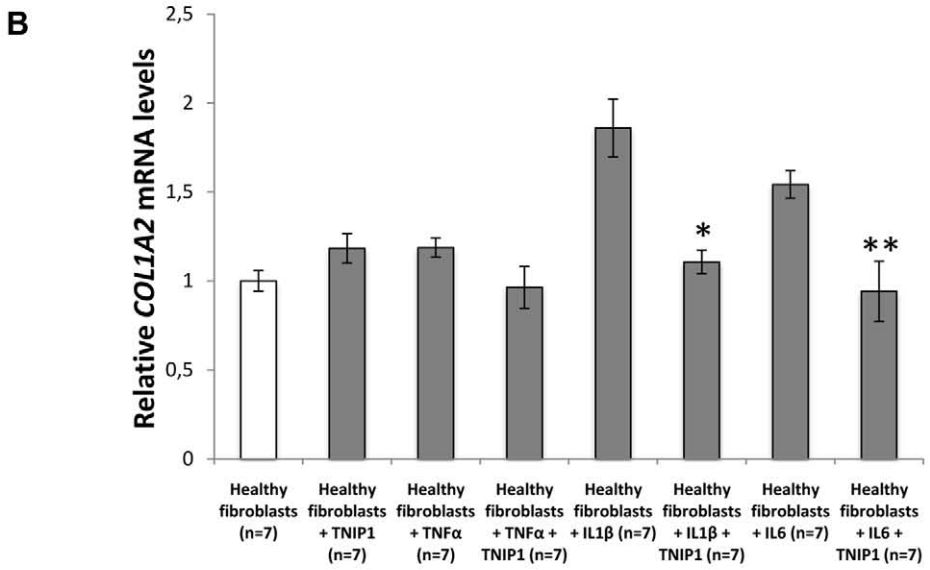
In conclusion, we have conducted a large genome-wide association study of SSc and identified two new SSc-risk loci, *PSORS1C1* and *TNIP1*. We also confirmed the association of SSc with variants at STAT4, IRF5 and CD247, in the European population. We also found compelling evidence of association to a putative new SSc risk locus on 2p24, close to the *RHOB* gene. None of the newly identified 3 loci have been previously reported associated to SSc. The *TNIP1* variants identified do not have precise functional implications; however, their localization within a regulatory region strongly suggests an impact on transcription of the gene. This is supported by our *ex vivo* and *in vitro* investigations. Altogether, our results are consistent with a reduced inhibition of NF-kappaB, therefore favoring inflammatory/immune responses and potentially contributing to the overproduction of extra-cellular matrix. This raises a new clue for a link between inflammation and SSc that could also be of importance in other fibrotic disorders.

Material and Methods

Study populations

Stage-1 included 654 SSc patients and 531 controls recruited through the French GENESYS project [11,13,30] and 2,003


A
Relative COL1A1 mRNA levels
2,5
2
1,5
1
0,5
0
*
**
Healthy fibroblasts (n=7)
Healthy fibroblasts + TNIP1 (n=7)
Healthy fibroblasts + TNFα (n=7)
Healthy fibroblasts + TNFα + TNIP1 (n=7)
Healthy fibroblasts + IL1β (n=7)
Healthy fibroblasts + IL1β + TNIP1 (n=7)
Healthy fibroblasts + IL6 (n=7)
Healthy fibroblasts + IL6 + TNIP1 (n=7)


B
Relative COL1A2 mRNA levels
2,5
2
1,5
1
0,5
0
*
**
Healthy fibroblasts (n=7)
Healthy fibroblasts + TNIP1 (n=7)
Healthy fibroblasts + TNFα (n=7)
Healthy fibroblasts + TNFα + TNIP1 (n=7)
Healthy fibroblasts + IL1β (n=7)
Healthy fibroblasts + IL1β + TNIP1 (n=7)
Healthy fibroblasts + IL6 (n=7)
Healthy fibroblasts + IL6 + TNIP1 (n=7)

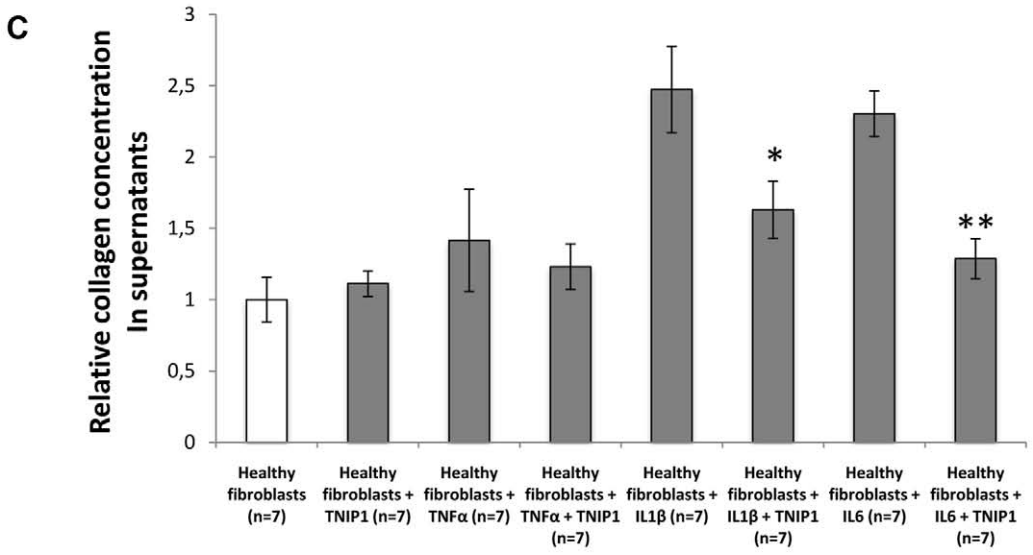

C
Relative collagen concentration In supernatants
3
2,5
2
1,5
1
0,5
0
*
**
Healthy fibroblasts (n=7)
Healthy fibroblasts + TNIP1 (n=7)
Healthy fibroblasts + TNFα (n=7)
Healthy fibroblasts + TNFα + TNIP1 (n=7)
Healthy fibroblasts + IL1β (n=7)
Healthy fibroblasts + IL1β + TNIP1 (n=7)
Healthy fibroblasts + IL6 (n=7)
Healthy fibroblasts + IL6 + TNIP1 (n=7)

Figure 5. TNIP1 abrogates the profibrotic effects of proinflammatory cytokines on collagen synthesis by healthy fibroblatsts. (A,B) Healthy dermal fibroblasts treated with recombinant Il1β or Il6 and incubated 24 hours with TNIP1 displayed decreased mRNA levels for (A) *COL1A1* (1.6 and 1.4-fold reduction, $P = 0.01$ and 0.03, respectively) and (B) *COL1A2* (1.7 and 1.6-fold reduction, $P = 0.02$ and 0.03, respectively). No significant effect was observed in cells treated with recombinant TNFα. (C) Collagen content in cell culture supernatants treated with Il1β or Il6 was also reduced upon treatment with TNIP1 (1.5 and 1.8-fold reduction, $P = 0.02$ and 0.03, respectively). * indicates a $P<0.05$ versus healthy control fibroblasts treated with recombinant IL1β. ** indicates a $P<0.05$ versus healthy control fibroblasts treated with recombinant IL6.
doi:10.1371/journal.pgen.1002091.g005

controls from the French Three-City (3C) cohort [21,22]. The stage-2 data included an independent collection of 4,492 samples (pre quality controls) from several University Hospitals in France, Italy, Germany and Eastern Europe. It also included 721 Italian controls recruited through nationwide efforts by HYPERGENE consortium and 481 Illumina HumanHap550 for the KORA S4 study [33], recruited in the city of Augsburg, Southern Germany. In both stage 1 and stage 2 samples, SSc patients fulfilled ACR criteria [34] and were classified in cutaneous subsets according to LeRoy's criteria [19]. Table 1 shows the main characteristics of the post-QC SSc patients and controls.

Ethics statement

All participants gave written informed consent, and approval was obtained from the relevant local ethical committees.

Genotyping and quality control analyses

Stage 1. French DNA samples from GENESYS and 3C were genotyped at Integragen and at the Centre National de Génotypage (Evry, France), respectively, with Illumina Human610-Quad BeadChip. Data were subjected to standard quality control procedures using tools implemented in PLINK version 1.07 [35]. Markers were removed if they had a genotype-missing rate >0.03 or a minor allele frequency (MAF)<0.05 or a Hardy-Weinberg $P<=10^{-5}$. Samples were removed on low (<98%) call rate, inconsistencies between reported gender and genotype-determined gender and/or genetic relatedness (identity-by-descent estimate >0.12). Applying these QC filters led to the removal of 791 subjects (56 cases, 735 controls). To detect individuals of non-European ancestry, we computed genome-wide average identity-by-state (IBS) distance with PLINK using a thinned map of 55,193 SNPs. To this end, we removed SNPs in extensive regions of LD (Chr.2, Chr.5, Chr.6, Chr.8, Chr.11) [36], and excluded SNPs if any pair within a 1000-SNPs window had $r^2>0.2$. Our data were then merged with genotypes at the same SNPs from 381 unrelated European (CEU), Yoruban (YRI) and Asian (CHB and JPT) samples from the HapMap project. Classical multi-dimensional scaling analysis was applied on the resulting matrix of IBS distances and the first two dimensions were extracted and plotted against each other. The HapMap data were clearly separated into three distinct clusters according to ancestry. Fifty seven of our stage-1 subjects did not cluster within the European group and were excluded from further analyses.

Stage 2. Follow-up analysis was conducted for the set of 17 SNPs that were identified in stage I and for 4 SNPs at STAT4, IRF5/TNPO3 and CD247 loci. *De novo* genotyping was performed by a competitive allele-specific PCR system (Kbioscience, Hoddeston, UK) [11,13,30]. The additional set of Italian and German control samples were previously genotyped using Illumina 1MQuad or Human610Quad bead chip. One SNP (rs9275224 in HLA) failed genotyping. Accuracy of genotyping was assessed using quality control procedures similar to those applied to our stage 1 data; following the quality control analyses, our stage 2 data consisted of 20 SNPs genotyped in a total of 1,682 cases and 3,926 controls (Table 1).

Statistical association analyses

Association analysis of the genotype data was conducted with PLINK (v1.07) software [35]. All reported P values are two sided. In stage 1, we applied logistic regression assuming an additive genetic model. The quantile-quantile plot was used to evaluate overall significance of the genome-wide association results and the potential impact of residual population substructure. A conservative genome-wide significance threshold of $0.05/489{,}918 = 1.02\times10^{-7}$ was used.

Stage 2 association and combined analyses were carried out with the Mantel-Haenszel test to control for differences between geographical groups. A Breslow-Day test was performed to assess the heterogeneity of effects in different populations. In the replication analysis, P values<0.05 and direction of effect as observed in the stage-1 data, were considered to indicate statistical significance.

Secondary statistical analyses were conducted to assess independency of multiple association signals within and between loci and homogeneity of effects between subgroups of SSc patients. Case-only association analyses were conducted using the three main clinical variables (Table 1). The LD structure of the identified loci was analyzed using Haploview 4.1 [37] and LD blocks delimited using the D′-based confidence interval method [38]. The locus-specific Population attributable risk (PAR) was calculated for each of the 6 replicated loci (HLA-DQB1, TNIP1, PSORS1C1, STAT4, IFR5/TNPO3 and CD247) according to the following formula: $PAR = RAF\times(OR-1)/(RAF\times(OR-1)+1)$, where RAF is the frequency of the associated allele in the controls, and OR is the odds ratio associated with the risk allele. The combined PAR was computed as $1-P_j(1-PAR_j)$.

Histologic and cytologic investigations

Fibroblast cultures were prepared by outgrowth cultures from lesional skin biopsy specimens of eleven SSc patients and from twelve healthy controls matched for age and sex. The median age of SSc patients was 49 years old (range: 22–67 years) and their median disease duration was 7 years (range: 1–17 years); seven had the limited cutaneous subset and four the diffuse. Immunohistochemistry was performed on paraffin-embedded skin sections from 5 SSc patients and 5 controls using mouse anti-human TNIP1 antibodies (eBioscience, Frankfurt, Germany). Total RNA, issued from cultured dermal fibroblasts, isolation and reverse transcription into complementary DNA were performed as previously described [39]. Gene expression was quantified by SYBR Green real-time PCR, with a specific primer pair available upon request. Protein assessment was performed on western blots, as previously described [40] using mouse anti-human TNIP1 antibodies (eBioscience, CA, USA). In selected experiments, dermal fibroblasts from healthy control subjects and patients with SSc were treated for 24 hours with recombinant TNIP1 (2 μg/ml, Abnova, Tapei City, Taiwan) in the presence or not of the following proinflammatory cytokines: TNFα (20 ng/ml, R&D systems, Abingdon, UK), IL1β (1 μg/ml, Immunotools, Friesoythe, Germany) or IL6 (1 μg/ml, Immunotools). mRNA levels of human α1(I) and α2(I) procollagen were quantified by quantitative

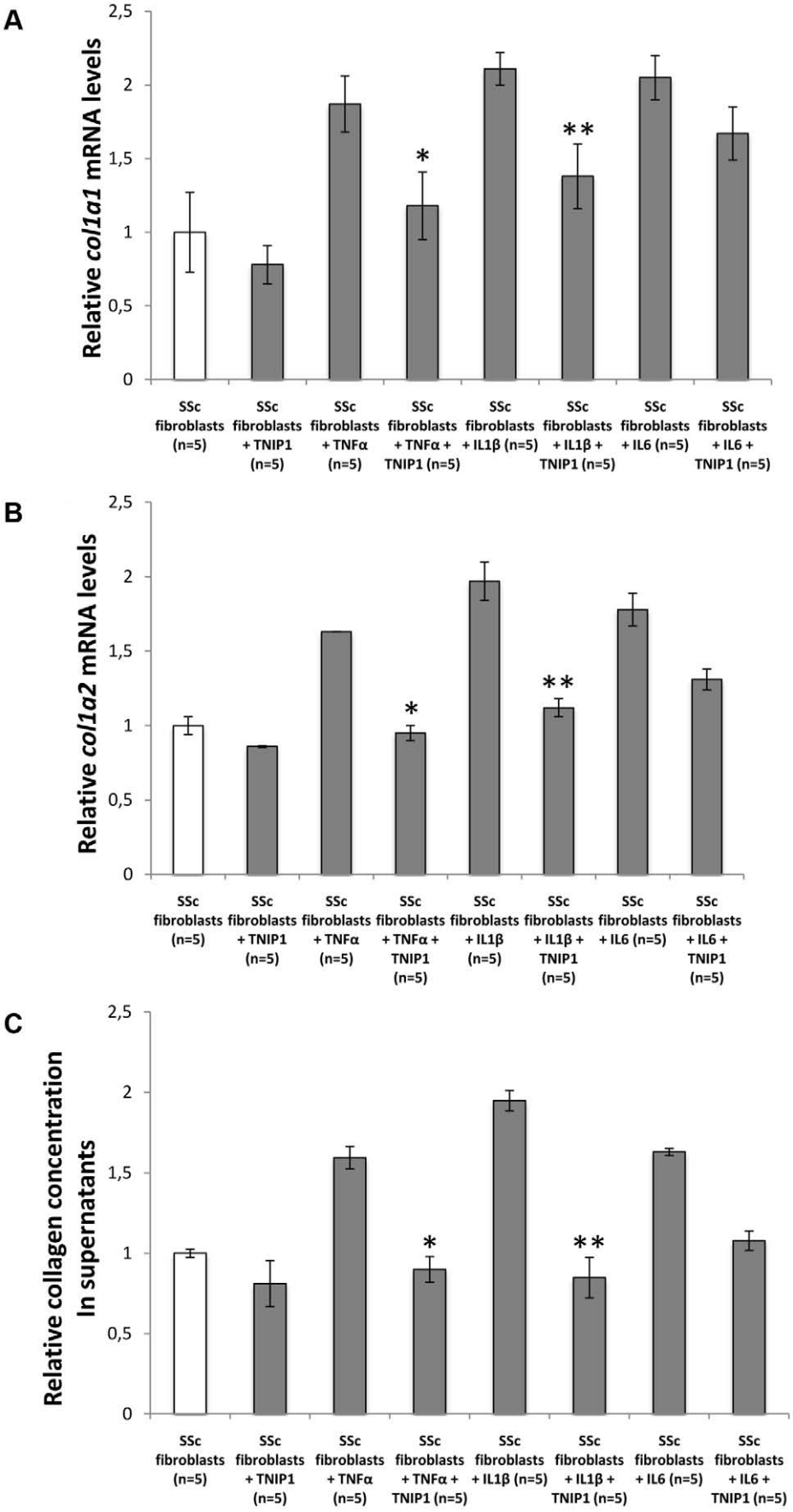

A
Relative col1a1 mRNA levels
2,5
2
1,5
1
0,5
0
*
**
SSc fibroblasts (n=5)
SSc fibroblasts + TNIP1 (n=5)
SSc fibroblasts + TNFα (n=5)
SSc fibroblasts + TNFα + TNIP1 (n=5)
SSc fibroblasts + IL1β (n=5)
SSc fibroblasts + IL1β + TNIP1 (n=5)
SSc fibroblasts + IL6 (n=5)
SSc fibroblasts + IL6 + TNIP1 (n=5)
B
Relative col1a2 mRNA levels
2,5
2
1,5
1
0,5
0
*
**
SSc fibroblasts (n=5)
SSc fibroblasts + TNIP1 (n=5)
SSc fibroblasts + TNFα (n=5)
SSc fibroblasts + TNFα + TNIP1 (n=5)
SSc fibroblasts + IL1β (n=5)
SSc fibroblasts + IL1β + TNIP1 (n=5)
SSc fibroblasts + IL6 (n=5)
SSc fibroblasts + IL6 + TNIP1 (n=5)
C
Relative collagen concentration In supernatants
2,5
2
1,5
1
0,5
0
*
**
SSc fibroblasts (n=5)
SSc fibroblasts + TNIP1 (n=5)
SSc fibroblasts + TNFα (n=5)
SSc fibroblasts + TNFα + TNIP1 (n=5)
SSc fibroblasts + IL1β (n=5)
SSc fibroblasts + IL1β + TNIP1 (n=5)
SSc fibroblasts + IL6 (n=5)
SSc fibroblasts + IL6 + TNIP1 (n=5)

Figure 6. TNIP1 abrogates the profibrotic effects of proinflammatory cytokines on collagen synthesis by SSc fibroblatsts. (A,B) SSc dermal fibroblasts treated with recombinant TNFa or IL1β and incubated 24 hours with TNIP1 displayed decreased mRNA levels for (A) *COL1A1* (1.6 and 1.5-fold reduction, $P = 0.02$ and 0.04, respectively) and (B) *COL1A2* (1.7 and 1.8-fold reduction, $P = 0.02$ and 0.009, respectively). No significant effect was observed in cells treated with recombinant IL6. (C) Collagen content in cell culture supernatants treated with TNFa or IL1β was also reduced upon treatment with TNIP1 (1.8 and 2.3-fold reduction, $P = 0.04$ and 0.03, respectively). * indicates a $P<0.05$ vs. SSc fibroblasts treated with recombinant TNFa. ** indicates a $P<0.05$ vs. SSc fibroblasts treated with recombinant IL1 β.
doi:10.1371/journal.pgen.1002091.g006

real-time PCR, specific primers are available upon request. The collagen content in cell culture supernatants was analyzed with the SirCol collagen assay (Biocolor, Belfast, UK) [41]. Comparisons were performed using Student's T test.

Supporting Information

Table S1 GWAS results for the most associated ($P<10^{-4}$) SNPs.
(DOC)

Table S2 Results of conditional logistic regression analysis for 28 SNPs ($P<10^{-3}$) in the MHC region.
(DOC)

Table S3 Results of Conditional logistic regression analysis for top 7 SNPs outside the MHC region in GWAS data.
(DOC)

Table S4 Association results in the combined (stage-1 and stage-2) data for the replicated SNPs by sub-type of SSc patients.
(DOC)

Acknowledgments

The EULAR Scleroderma Trials and Research group (EUSTAR) facilitated the DNA collection.

Centers that contributed to the recruitment of SSc patients: Carpentier PH (Grenoble, France), Sibilia J (Strasbourg, France), Amoura Z, Frances C (Paris, France), Cosnes A (Créteil, France), and Bichat Paris CRB (Dr Benessiano and Pr Grandchamp) for the controls.

Author Contributions

Conceived and designed the experiments: Y Allanore, M Saad, P Dieudé, J Avouac, JHW Distler, C Boileau, M Martinez. Performed the experiments: Y Allanore, M Saad, P Dieudé, J Avouac, C Boileau, M Martinez. Analyzed the data: Y Allanore, M Saad, P Dieudé, J Avouac, JHW Distler, C Boileau, M Martinez. Wrote the paper: Y Allanore, M Saad, P Dieudé, J Avouac, C Boileau, M Martinez. Revision of the manuscript: Y Allanore, M Saad, P Dieudé, J Avouac, JHW Distler, P Amouyel, M Matucci-Cerinic, G Riemekasten, P Airo, I Melchers, E Hachulla, D Cusi, H-E Wichmann, J Wipff, J-C Lambert, N Hunzelmann, K Tiev, P Caramaschi, E Diot, O Kowal-Bielecka, G Valentini, L Mouthon, L Czirják, N Damjanov, E Salvi, C Conti, M Müller, U Müller-Ladner, V Riccieri, B Ruiz, J-L Cracowski, L Letenneur, A-M Dupuy, O Meyer, A Kahan, A Munnich, C Boileau, M Martinez. Contribution to sampling (biomaterial and clinical data): Y Allanore, P Dieudé, J Avouac, JHW Distler, P Amouyel, M Matucci-Cerinic, G Riemekasten, P Airo, I Melchers, E Hachulla, D Cusi, H-E Wichmann, J Wipff, J-C Lambert, N Hunzelmann, K Tiev, P Caramaschi, E Diot, O Kowal-Bielecka, G Valentini, L Mouthon, L Czirják, N Damjanov, E Salvi, C Conti, M Müller, U Müller-Ladner, V Riccieri, B Ruiz, J-L Cracowski, L Letenneur, A-M Dupuy, O Meyer, A Kahan, A Munnich.

References

1. Valentini G, Black C (2002) Systemic sclerosis. Best Pract Res Clin Rheumatol 16: 807–16.
2. Thompson AE, Pope JE. Increased prevalence of scleroderma in southwestern Ontario: a cluster analysis. J Rheumatol 29: 1867–73.
3. Le Guern V, Mahr A, Mouthon L, Jeanneret D, Carzon M, et al. (2004) Prevalence of systemic sclerosis in a French multi-ethnic county. Rheumatology (Oxford) 2004;43: 1129–37.
4. Czirjak L, Kiss CG, Lovei C, Süto G, Varjú C, et al. (2005) Survey of Raynaud's phenomenon and systemic sclerosis based on a representative study of 10,000 south-Transdanubian Hungarian inhabitants. Clin Exp Rheumatol 23: 801–8.
5. Arnett FC, Cho M, Chatterjee S, Aguilar MB, Reveille JD, et al. (2001) Familial occurrence frequencies and relative risks for systemic sclerosis (scleroderma) in three United States cohorts. Arthritis Rheum 44: 1359–62.
6. Feghali-Bostwick C, Medsger TA, Jr., Wright TM (2003) Analysis of systemic sclerosis in twins reveals low concordance for disease and high concordance for the presence of antinuclear antibodies. Arthritis Rheum 48: 1956–6.
7. Gabrielli A, Avvedimento EV, Krieg T (2009) Scleroderma. N Engl J Med 360: 1989–2003.
8. Lafyatis R, York M (2009) Innate immunity and inflammation in systemic sclerosis. Curr Opin Rheumatol 21: 617–22.
9. Varga J, Abraham D (2007) Systemic sclerosis: a prototypic multisystem fibrotic disorder. J Clin Invest 117: 557–67.
10. Allanore Y, Dieude P, Boileau C (2010) Updating the genetics of Systemic Sclerosis. Curr Opin Rheumatol 22: 665–70.
11. Dieudé P, Guedj M, Wipff J, Avouac J, Fajardy I, et al. (2009) Association between the IRF5 rs2004640 functional polymorphism and systemic sclerosis: a new perspective for pulmonary fibrosis. Arthritis Rheum 60: 225–33.
12. Ito I, Kawaguchi Y, Kawasaki A, Hasegawa M, Ohashi J, et al. (2009) Association of a functional polymorphism in the IRF5 region with systemic sclerosis in a Japanese population. Arthritis Rheum 60: 1845–50.
13. Dieudé P, Guedj M, Wipff J, Ruiz B, Hachulla E, et al. (2009) STAT4 is a genetic risk factor for systemic sclerosis having additive effects with IRF5 on disease susceptibility and related pulmonary fibrosis. Arthritis Rheum 60: 2472–9.
14. Rueda B, Broen J, Simeon C, Hesselstrand R, Diaz B, et al. (2009) The STAT4 gene influences the genetic predisposition to systemic sclerosis phenotype. Hum Mol Genet 18: 2071–7.
15. Gourh P, Agarwal SK, Divecha D, Assassi S, Paz G, et al. (2009) Polymorphisms in TBX21 and STAT4 increase the risk of systemic sclerosis: evidence of possible gene-gene interaction and alterations in Th1/Th2 cytokines. Arthritis Rheum 60: 3794–806.
16. Arnett FC, Gourh P, Shete S, Ahn CW, Honey RE, et al. (2010) Major histocompatibility complex (MHC) class II alleles, haplotypes, and epitopes which confer susceptibility or protection in the fibrosing autoimmune disease systemic sclerosis: analyses in 1300 Caucasian, African-American and Hispanic cases and 1000 controls. Ann Rheum Dis 69: 822–7.
17. Zhou X, Lee JE, Arnett FC, Xiong M, Park MY, et al. (2009) HLA-DPB1 and DPB2 are genetic loci for systemic sclerosis: a genome-wide association study in Koreans with replication in North Americans. Arthritis Rheum 60: 3807–14.
18. Radstake TR, Gorlova O, Rueda B, Martin JE, Alizadeh BZ, et al. (2010) Genomewide association study of systemic sclerosis identifies CD247 as a new susceptibility locus. Nat Genet 42: 426–9.
19. LeRoy EC, Black C, Fleischmajer R, Jablonska S, Krieg T, et al. (1988) Scleroderma (systemic sclerosis): classification, subsets and pathogenesis. J Rheumatol 15: 202–205.
20. Wollheim FA (2005) Classification of systemic sclerosis. Visions and reality. Rheumatology 44: 1212–6.
21. Lambert JC, Heath S, Even G, Campion D, Sleegers K, et al. (2009) Genomewide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat Genet 41: 1094–1099.
22. C Study Group (2003) Vascular factors and risk of dementia: design of the Three-City Study and baseline characteristics of the study population. Neuroepidemiology 22: 316–25.
23. Fan X, Yang S, Huang W, Wang ZM, Sun LD, et al. (2008) Fine mapping of the psoriasis susceptibility locus PSORS1 supports HLA-C as the susceptibility gene in the Han Chinese population. PLoS Genet 4: e1000038. doi:10.1371/journal.pgen.1000038.
24. Reich K, Hüffmeier U, König IR, Lascorz J, Lohmann J, et al. (2007) TNF polymorphisms in psoriasis: association of psoriatic arthritis with the promoter polymorphism TNF*-857 independent of the PSORS1 risk allele. Arthritis Rheum 56: 2056–64.
25. Wertz IE, O'Rourke KM, Zhou H, Eby M, Aravind L, et al. (2004) De-ubiquitination and ubiquitin ligase domains of A20 downregulate NF-kappaB signalling. Nature 430: 694–699.
26. Boone DL, Turer EE, Lee EG, Ahmad RC, Wheeler MT, et al. (2004) The ubiquitin-modifying enzyme A20 is required for termination of Toll-like receptor responses. Nat Immunol 5: 1052–1060.

27. Gateva V, Sandling JK, Hom G, Taylor KE, Chung SA, et al. (2009) A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus. 41: 228–233.
28. Han JW, Zheng HF, Cui Y, Sun LD, Ye DQ, et al. (2009) Genomewide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. Nat Genet 41: 1234–1237.
29. Nair RP, Duffin KC, Helms C, Ding J, Stuart PE, et al. (2009) Genomewide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. Nat Genet 41: 199–204.
30. Dieudé P, Guedj M, Wipff J, Ruiz B, Riemekasten G, et al. (2010) Association of the TNFAIP3 rs5029939 variant with systemic sclerosis in the European Caucasian population. Ann Rheum Dis 69: 1958–64.
31. Falgarone G, Chiocchia G (2009) Clusterin: A multifacet protein at the crossroad of inflammation and autoimmunity. Adv Cancer Res 104: 139–70.
32. Gabrielli A, et al. (2007) Stimulatory autoantibodies to the PDGF receptor: a link to fibrosis in scleroderma and a pathway for novel therapeutic targets. Autoimmun Rev 7: 121–6.
33. Wichmann HE, Gieger C, Illig T (2005) KORA-gen-resource for population genetics, controls and a broad spectrum of disease phenotypes. Gesundheitswesen 67, Suppl. 1: S26–S30.
34. Anonymous (1980) Subcommittee for Scleroderma Criteria of the American Rheumatism Association Diagnostic and Therapeutic Criteria Committee: Preliminary criteria for the classification of systemic sclerosis (scleroderma). Arthritis Rheum 23: 581–90.
35. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575.
36. Price AL, Weale ME, Patterson N, Myers SR, Need AC, et al. (2008) Long-range LD can confound genome scans in admixed populations. Am J Hum Genet 83: 132–135.
37. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263–265.
38. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, et al. (2002) The structure of haplotype blocks in the human genome. Science 296: 2225–2229.
39. Distler JH, Jüngel A, Huber LC, Schulze-Horsel U, Zwerina J, et al. (2007) Imatinib mesylate reduces production of extracellular matrix and prevents development of experimental dermal fibrosis. Arthritis Rheum 56: 311–322.
40. Avouac J, Wipff J, Goldman O, Ruiz B, Couraud PO, et al. (2008) Angiogenesis in systemic sclerosis: impaired expression of vascular endothelial growth factor receptor 1 in endothelial progenitor-derived cells under hypoxic conditions. Arthritis Rheum 58: 3550–3561.
41. Reich N, Maurer B, Akhmetshina A, Venalis P, Dees C, et al. (2010) The transcription factor Fra-2 regulates the production of extracellular matrix in systemic sclerosis. Arthritis Rheum 62: 280–290.

Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies

*International Parkinson Disease Genomics Consortium**

Summary

Background Genome-wide association studies (GWAS) for Parkinson's disease have linked two loci (*MAPT* and *SNCA*) to risk of Parkinson's disease. We aimed to identify novel risk loci for Parkinson's disease.

Methods We did a meta-analysis of datasets from five Parkinson's disease GWAS from the USA and Europe to identify loci associated with Parkinson's disease (discovery phase). We then did replication analyses of significantly associated loci in an independent sample series. Estimates of population-attributable risk were calculated from estimates from the discovery and replication phases combined, and risk-profile estimates for loci identified in the discovery phase were calculated.

Findings The discovery phase consisted of 5333 case and 12 019 control samples, with genotyped and imputed data at 7 689 524 SNPs. The replication phase consisted of 7053 case and 9007 control samples. We identified 11 loci that surpassed the threshold for genome-wide significance ($p<5\times10^{-8}$). Six were previously identified loci (*MAPT, SNCA, HLA-DRB5, BST1, GAK* and *LRRK2*) and five were newly identified loci (*ACMSD, STK39, MCCC1/LAMP3, SYT11*, and *CCDC62/HIP1R*). The combined population-attributable risk was 60·3% (95% CI 43·7–69·3). In the risk-profile analysis, the odds ratio in the highest quintile of disease risk was 2·51 (95% CI 2·23–2·83) compared with 1·00 in the lowest quintile of disease risk.

Interpretation These data provide an insight into the genetics of Parkinson's disease and the molecular cause of the disease and could provide future targets for therapies.

Funding Wellcome Trust, National Institute on Aging, and US Department of Defense.

Published Online February 2, 2011 DOI:10.1016/S0140-6736(10)62345-8

See Online/Comment DOI:10.1016/S0140-6736(11)60062-7

*Members listed at end of paper

Correspondence to: Dr Andrew Singleton, Laboratory of Neurogenetics, 35 Convent Drive, Bethesda, MD 20837, USA **singleta@mail.nih.gov**

Introduction

Parkinson's disease was long thought to be a non-genetic disease. Recent advances in genotyping have enabled large-scale assessment of genetic risk factors associated with Parkinson's disease. *MAPT* and *SNCA*[1–5] have consistently shown associations in genome-wide association studies (GWAS). *BST1, LRRK2, GAK,* and *HLA-DRB5*[1–3,5–7] have been implicated in some studies but these associations have not been definitively confirmed.

Although an exciting next step in the genetic study of human disease will be the use of exome or genome sequencing in adequately powered large-scale population-based studies, this method is cost prohibitive at present.[8] A compromise between array-based and sequence-based methods is the use of freely available sequence-based resources from the 1000 Genomes Project, which allows imputation of a large number of variants into existing genotyping studies.

We did a meta-analysis of Parkinson's disease GWAS to investigate the associations of previously identified loci and identify novel risk loci for Parkinson's disease.

Methods

Study design

Investigators representing four published GWAS[1–3,9] formed a consortium with the predetermined goal to discover new loci associated with Parkinson's disease by a prospective meta-analysis of imputed sequence variants (discovery stage). We identified one additional dataset from the database of genotypes and phenotypes.[5,10] A secondary requirement for inclusion of a dataset in this study was the ability to use custom-built ImmunoChip arrays to do replication analyses in independent sample series (replication stage). The five included datasets were from the USA National Institute on Aging, UK, Germany, France, and the USA database of genotypes and phenotypes.[1–3,5,9,11]

We aimed to assess the biological consequences of risk variants identified in our study by examining the association between these alleles and both gene expression and DNA methylation. Our primary interest in these single nucleotide polymorphism (SNP)-based analyses was to investigate every locus associated with Parkinson's disease and to test whether the disease-related SNPs were associated with DNA methylation or gene expression levels. Further, we wanted to test whether the most strongly disease-associated SNPs were also the most strongly associated quantitative trait locus SNPs.

For the **1000 Genomes Project** see http://www.1000genomes.org/

Data imputation and statistical analysis

After individual sample collection and study-specific quality control (webappendix pp 1–7), we used a Markov

See Online for webappendix

Chain based haplotyper (MACH; version 1.0.16) on every dataset to impute genotypes for all participants of European ancestry with haplotypes derived from initial low coverage sequencing of 112 European ancestry samples in the 1000 Genomes Project (as of August, 2009).[12,13] For all datasets, data were imputed by a two-stage design. The first stage generated error and crossover maps as parameter estimates for imputation on a random subset of 200 samples per study over 100 iterations of the initial statistical model. We used these parameter estimates to generate maximum likelihood estimates of allele numbers per SNP on the basis of reference haplotypes for the datasets during the second stage of the imputation. SNPs with RSQR quality estimates of less than 0·30 as indicated by MACH were excluded from analyses of the datasets, because imputed genotypes below this threshold are probably of poor quality.

We did genome-wide dataset analyses at every site with MACH2DAT.[12] We used non-integer allele numbers as a primary predictor of Parkinson's disease in logistic regression models to account for imputation uncertainty. Webappendix pp 1–7 shows specific details of analyses of the datasets. Summary statistics from genome-wide association analyses of every dataset were included in the meta-analyses. For every dataset, we used basic covariates of component vectors 1 and 2 from either principal components or multidimensional scaling analyses of the case-control cohorts to identify random genomic differences between genotyped data from cases and controls in the discovery phase, which were used to adjust statistical models for covariates accounting for possible population substructure. This adjustment was not done in analyses of the UK dataset in the discovery phase of analyses.

For the replication step, we included the SNPs that passed genome-wide significance (fixed effects $p<1\times10^{-5}$) and quality control on a custom ImmunoChip array (Illumina, San Diego, CA, USA) in collaboration with the Sanger Institute (Hinxton, UK). Additionally, we analysed two GWAS (from the Netherlands and Iceland) after the meta-analysis and included these in the replication stage by the same imputation procedure. These data were provided by consortium members who provided the GWAS data after the initial discovery phase. We included a quality control step in the replication analyses that removed SNPs with inconsistent results across the datasets (I^2>75%).[14,15] Webappendix pp 4–8 shows detailed descriptions of the replication analyses that were done in the five ImmunoChip replication cohorts (USA, UK, Netherlands, Germany, and France) and two in-silico GWAS datasets (Iceland and Netherlands).

We did fixed-effects inverse variance-weighted meta-analyses with meta-analysis helper (METAL),[16,17] with the standard errors of the β coefficients scaled by the square root of study-specific genomic inflation factor estimates before combining the summary statistics across datasets. We calculated genomic control for both individual datasets and the entire meta-analysis for quality control. Genomic control is often estimated as the deviance of the median test statistic distribution from the expected null; genomic inflation factors less than 1·05 are the general standard in GWAS.[18] We used fixed-effects meta-analyses as the primary method of discovery and R (version 2.11) to do a secondary random-effects meta-analysis for every SNP.[19] This second analysis is useful to estimate the possible effect of study heterogeneity on results and to qualitatively infer the effect of study heterogeneity on replication success and generalisability for similar sample series. We calculated χ^2 tests for heterogeneity (Cochran's Q) with METAL and we generated I^2 estimates with R. Meta-analyses and estimates of study heterogeneity were re-run with PLINK as a quality control measure.

We calculated risk-profile estimates on the basis of cumulative load of risk alleles for loci identified in the discovery phase, weighted by the discovery phase effect estimates ($\log_{\text{odds ratio}}$). This profile model was applied to the ImmunoChip genotyped replication cohorts, and the effects were combined across cohorts by inverse variance weighting.

Population-attributable risk was estimated for the specific genetic contribution to disease of the risk loci

	Cases			Controls			Genome-wide association study			
	Sample size	Women (%)	Mean age at onset (years [SD])	Sample size	Women (%)	Mean age at examination (years [SD])	Number of successfully genotyped SNPs	Number of successfully imputed SNPs	Covariates	Genomic inflation factor (λ)
USA–NIA	971	40·5%	55·9 (15·1)	3034	52·8%	62 (15·6)	463 187	7 590 773	Population structure*	1·043
UK	1705	43·3%	65·8 (10·8)	5200	49·5%	NA	598 821	7 678 643	None	1·048
Germany	742	39·8%	56 (11·6)	944	48·0%	NA	463 187	7 589 890	Population structure*	1·029
France	1039	41·2%	48·9 (12·8)	1984	33·0%	73·7 (5·4)	493 081	7 340 040	Population structure*	1·029
USA–dbGAP	876	40·4%	61·5 (9·2)	857	60·2%	NA	334 513	7 482 040	Population structure*	1·023
Meta-analysis	5333	..	..	12 019	..	..	547 951	7 689 524†	Genomic control‡	1·042

NIA=National Institute on Aging. NA=not available. dbGAP=database of genotypes and phenotypes. *Adjusted for component vectors 1 and 2 from multidimensional scaling analyses of the study population. †Passed quality control in at least two of the included datasets; ‡Summary statistics scaled by study-specific genomic inflation factors before meta-analysis.

Table 1: **Dataset characteristics**

www.thelancet.com **Published online February 2, 2011 DOI:10.1016/S0140-6736(10)62345-8**

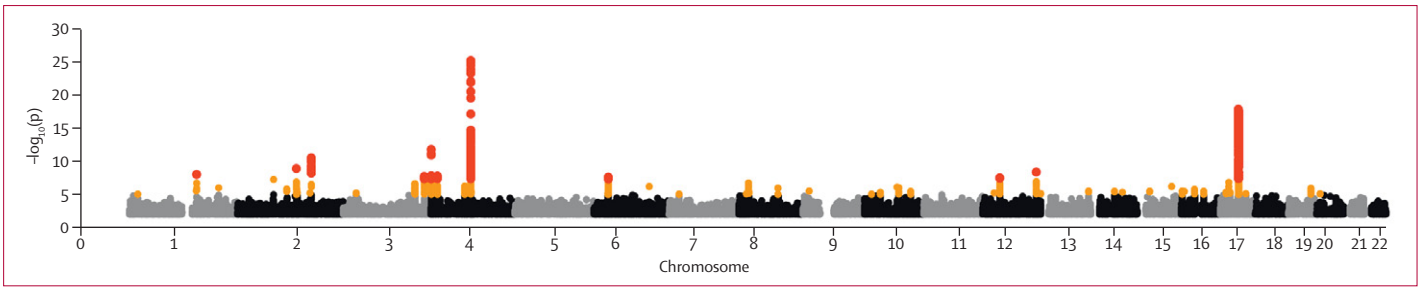


Figure 1: **Manhattan plot of Parkinson's disease associations for all SNPs in the discovery phase**
p values from fixed-effects meta-analysis for 7 689 524 SNPs successfully imputed or genotyped in at least two individual datasets. Genomic inflation factor=1·035. Red points=SNPs with $p<5\times10^{-8}$. Orange points=SNPs with p values ranging from less than 1×10^{-5} to 5×10^{-8}. Regions containing red points were followed up in replication analyses. SNP=single nucleotide polymorphism.

identified. In broad terms, it estimates the decrease in cases of a particular disease within a population that would occur if the risk factor were removed from that population. Effect sizes and minor allele frequencies were calculated from joint estimates from the discovery and replication phases combined, to lessen the overestimation caused by the so-called winner's curse—a form of ascertainment bias that often occurs in two-stage GWAS wherein natural genetic variation contributes to a slight overestimation of effect sizes in the discovery phase.[20] Joint estimates were also used because of their large sample size, which should generate more accurate effect estimates.

DNA methylation values at sites close to the risk variants and the expression of genes within the risk loci were treated as quantitative traits, and we assessed whether the alleles of SNPs across the risk loci were associated with either, denoting a quantitative trait locus. Webappendix pp 10–13 and a previous study[21] provide more detail about the methods used to map quantitative trait loci. Briefly, we used dense genotype data generated in up to 350 people who had donated brain tissue and were neurologically healthy at the time of death. DNA methylation and transcript expression were assessed in frozen tissue from both frontal cortex and cerebellum samples of every brain. We tested the association between any typed polymorphism and any assayed DNA methylation site or transcript. All SNPs within 1 mb from the SNPs with the smallest p values per locus with fixed-effects p values less than 1×10^{-5} were investigated as candidate loci that affect the expression and methylation values of proximal mRNA transcript probes and CpG methylation sites. With the minor allele as a reference for directionality, we used linear models to quantify the relation between quantitative trait loci and risk effect for all the loci that contained significant quantitative trait locus associations. We used linkage-adjusted Bonferroni correction for significance (webappendix pp 11–13).

Role of the funding source

The sponsors of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. All members of the writing group had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

The discovery phase included 5333 case and 12 019 control samples, with genotyped and imputed data at 7 689 524 SNPs (table 1; figure 1). 7053 case and 9007 control samples were included in analyses in the replication stage. Results of tests across the software packages METAL, R, and PLINK differed only slightly (data not shown).

We identified 11 loci that surpassed the threshold for genome-wide significance ($p<5\times10^{-8}$; table 2).[18] One locus on chromosome 17 from 18 601 523 to 18 602 998 bp that contained six SNPs associated with Parkinson's disease in the UK cohort was not included because p values were not less than 0·1 in any other cohort and I^2 was greater than 75%. For simplicity, we have focussed only on the most significant SNP per locus that met these criteria, and its nearest gene or genes. However, we do recognise that the most proximal gene is not necessarily the gene functionally affected by risk alleles and that GWAS identify loci rather than specific genes. Webappendix pp 15–31 provides more detailed results for every region.

We confirmed the Parkinson's disease associations at the *SNCA* and *MAPT* loci[1–5] and at *BST1, LRRK2, GAK,* and *HLA-DRB5* (table 2).[1–3,5–7] Additionally, we detected evidence of associations at five new loci (the closest gene to the top SNP at every loci is *ACMSD, STK39, MCCC1/LAMP3, SYT11,* and *CCDC62/HIP1R*). However, two of these loci (*ACMSD* and *CCDC62/HIP1R*) showed moderate evidence of heterogeneity across populations. Webappendix pp 7–8 provides further details about the replication phase of analysis. All five novel loci passed a Bonferroni threshold of $p<0·0045$ (correcting for the 11 SNPs tested in the replication phase) for association in the replication phase.

The SNP at the *SYT11* locus is about 650 kb from the known Parkinson's disease risk factor gene *GBA* and its pseudogene, in a region of the genome with low recombination.[22] To test whether this proximity might contribute to the possible co-segregation of risk alleles at the *SYT11* locus and *GBA* risk mutations we analysed data from a subset of patients with Parkinson's disease

	C	Position (bp)	MAF in discovery phase	Minor/major alleles	Candidate gene	Discovery phase					Replication phase		Combined PAR estimate (95% CI)
						OR (SE) per minor allele dose	Fixed effects p value	Random effects p value	I^2 index (%)	I^2 p value	OR (SE) per minor allele dose	Fixed effects p value	
chr1:154105678	1	154105678	0·02	T/C	*SYT11*	1·67 (0·09)	$1·02\times10^{-8}$	$5·70\times10^{-9}$	0·00%	0·77	1·44 (0·08)	$1·18\times10^{-6}$	1·21% (0·34–1·47)
rs6710823	2	135308851	0·19	A/G	*ACMSD*	1·38 (0·05)	$1·35\times10^{-9}$	$1·61\times10^{-5}$	48·26%	0·11	1·07 (0·02)	0·003161	4·05% (1·66–6·82)
rs2102808	2	168825271	0·13	T/G	*STK39*	1·28 (0·04)	$3·31\times10^{-11}$	$1·54\times10^{-11}$	0·00%	0·72	1·12 (0·04)	0·001639	2·29% (1·11–2·98)
rs11711441	3	184303969	0·14	A/G	*MCCC1/LAMP3*	0·82 (0·04)	$2·10\times10^{-8}$	$1·17\times10^{-8}$	0·00%	0·97	0·87 (0·03)	$6·92\times10^{-5}$	13·71% (9·05–17·70)
chr4:911311	4	911311	0·28	C/G	*GAK*	1·21 (0·03)	$1·80\times10^{-12}$	$2·96\times10^{-7}$	51·58%	0·09	1·14 (0·02)	$7·46\times10^{-8}$	4·87% (2·68–6·38)
rs11724635	4	15346199	0·45	C/A	*BST1*	0·87 (0·03)	$1·85\times10^{-8}$	0·001407	74·77%	$4·1\times10^{-3}$	0·87 (0·02)	$2·43\times10^{-9}$	7·82% (5·30–9·47)
rs356219	4	90856624	0·39	G/A	*SNCA*	1·30 (0·03)	$7·90\times10^{-26}$	$1·11\times10^{-26}$	0·00%	0·58	1·27 (0·02)	$4·23\times10^{-23}$	9·71% (6·68–10·27)
chr6:32588205	6	32588205	0·15	G/A	*HLA-DRB5*	0·70 (0·06)	$2·58\times10^{-8}$	$1·44\times10^{-8}$	0·00%	0·88	0·80 (0·04)	$9·30\times10^{-8}$	17·68% (11·04–23·00)
rs1491942	12	38907075	0·21	G/C	*LRRK2*	1·19 (0·03)	$3·23\times10^{-8}$	$5·24\times10^{-6}$	35·52%	0·20	1·30 (0·05)	$1·06\times10^{-8}$	2·09% (1·00–2·50)
rs12817488	12	121862247	0·46	A/G	*CCDC62/HIP1R*	1·16 (0·03)	$4·43\times10^{-9}$	$2·99\times10^{-6}$	34·97%	0·20	1·13 (0·03)	$9·06\times10^{-7}$	5·56% (3·20–7·37)
rs2942168	17	41070633	0·22	A/G	*MAPT*	0·76 (0·03)	$1·62\times10^{-18}$	$3·91\times10^{-19}$	0·00%	0·74	0·80 (0·03)	$1·37\times10^{-13}$	17·57% (12·92–20·78)

Only loci with $p<5\times10^{-8}$ in the meta-analysis are shown. The SNP with the smallest p value per locus on the basis of a fixed effects meta-analysis is shown. Webappendix pp 15–31 provide additional details for the associated loci described above. An expanded version of this table that shows all p values less than 1×10^{-5} from the discovery phase of analyses is available upon request. C=chromosome. MAF=minor allele frequency. OR=odds ratio. PAR=population-attributable risk. I^2 index=I^2 index of heterogeneity. I^2 p value=heterogeneity p value.

Table 2: **Summary of significant loci**

	p value	AUC	Risk quintile OR (95%CI)				
			First (reference group)	Second	Third	Fourth	Fifth
USA	$<2\times10^{-16}$	0·584	1·00	1·49 (1·25–1·78)	1·67 (1·40–2·00)	1·90 (1·59–2·27)	2·25 (1·88–2·70)
UK	$<2\times10^{-16}$	0·631	1·00	1·63 (1·27–2·08)	2·26 (1·77–2·88)	2·65 (2·09–3·38)	3·30 (2·60–4·21)
Germany	$1·44\times10^{-8}$	0·69	1·00	1·16 (0·86–1·57)	1·55 (1·14–2·11)	1·68 (1·23–2·29)	2·06 (1·51–2·82)
France	$6·15\times10^{-9}$	0·644	1·00	1·24 (0·72–2·16)	2·13 (1·26–3·66)	2·84 (1·68–4·88)	4·31 (2·51–7·55)
Netherlands	$8·34\times10^{-4}$	0·576	1·00	1·21 (0·74–2·00)	1·12 (0·68–1·84)	1·50 (0·93–2·42)	1·89 (1·17–3·07)
Combined	$<2\times10^{-16}$	0·63	1·00	1·43 (1·27–1·62)	1·77(1·55–1·99)	2·03 (1·80–2·32)	2·51 (2·23–2·83)
Cases (%)	..	..	886 (39·00%)	1069 (47·13%)	1185 (52·16%)	1268 (55·93%)	1394 (61·17%)

Combined analyses showed low heterogeneity of effect (Cochran's *Q* p>0·01). AUC=area under curve, indicated by the *c* index from receiver operator curves. OR=odds ratio.

Table 3: **Summary of risk-profile analyses**

who were included in the discovery phase analysis and who were from the USA, France, and Germany, in whom carriers of the *GBA* mutation have been identified. The results of this analysis suggested that the signal at the *SYT11* locus is independent of *GBA* mutations (webappendix pp 9–10 and pp 32–33). Similarly, we did an analysis that controlled for the known common *LRRK2* mutation G2019S in the replication phase, and showed that the association detected in the meta-analysis close to *LRRK2* might be caused by variation independent of this mutation (webappendix pp 9–10 and pp 32–33).

Table 2 shows the combined population-attributable risk estimates. The combined estimate across all 11 identified loci was 60·3% (95% CI 43·7–69·3). For the *MAPT* and *SNCA* loci alone it was 25·6% (18·7–28·9), which was higher than the previous estimate of about 20%.[2] The additional loci identified in this study (*ACMSD, STK39, MCCC1/LAMP3,* and *CCDC62/HIP1R*) had a combined estimate of 46·7% (30·7–56·8).

The odds ratio was 2·5 times higher in the highest quintile of disease risk than in the lowest quintile of disease risk (table 3). The *c* index from receiver operator curves in the pooled cohorts was 0·63.

We identified quantitative trait associations at 18 969 SNPs spread across five of the identified Parkinson's disease risk loci (summarised in figure 2, with complete results available from the authors upon request). The *MAPT* locus had many such associations, with 95·9% of all associations detected across all tissues and arrays in this region. For the *MAPT* locus, risk estimates were positively associated with quantitative trait locus effects, leading to increased gene expression (cerebellum r^2 0·1366; frontal cortex r^2 0·8042; both tissues $p<2\times10^{-16}$). We noted the opposite effect in the methylation quantitative trait loci at *MAPT*, with minor

alleles associated with increased risk and with decreased methylation (cerebellum r^2 0·9268, p <2×10^{-16}; frontal cortex r^2 0·4667, p=3·68×10^{-6}). The *MAPT* locus showed associations for multiple probes within *MAPT* and probes within proximal genes, including *ARL17A* and *PLEKHM1*. Methylation quantitative trait loci in the *MAPT* region included probes within *KIAA1267*, *LRRC37A*, and *NSF*.

Two SNPs in the *ACMSD* locus showed substantial associations, with expression levels in cerebellar tissues denoted by a proximal expression probe against the transcript *MCM6*, whose transcription start site is more than 750 kb from either associated SNP; the minor alleles at this pair of SNPs are associated with increased risk of Parkinson's disease and decreased gene expression. In samples from the frontal cortex, DNA methylation values at one CpG site within *FGFRL1* were associated with 27 proximal SNPs in the *GAK* region (>20 kb from the nearest SNP associated with Parkinson's disease), and all effect estimates suggested risk of Parkinson's disease and increased methylation (r^2 0·9897, p<2×10^{-16}).

The *HLA-DRB5* region contained 729 significant quantitative trait locus associations in the frontal and cerebellar tissue samples. For *HLA-DRB5*, we recorded an overall result similar to that identified at *MAPT*, with minor alleles associated with risk effects and with decreased DNA methylation (cerebellum r^2 0·4037, p=9·67×10^{-5}; frontal cortex r^2 0·4977, p<2×10^{-16}).

In addition to probes within *HLA-DRB5*, methylation quantitative trait locus associations were detected within probes tagging CpG sites in *BTNL2*, *HLA-DQB2*, and *SLC44A4*. One CpG probe in the cerebellum samples was associated with 18 SNPs in the *CCDC62/HIP1R* region. We noted a relation between risk alleles and DNA methylation at a CpG site within *GPR109B*, for which increased risk estimates were closely associated with more negative methylation effects (cerebellum r^2 0·4977, p<2×10^{-16}).

Both the *LRRK2* and *SNCA* genes play a part in Parkinson's disease; thus, we examined these loci further for potential quantitative trait loci. Detection of *LRRK2* with the array-based method showed expression that was too low to do an accurate analysis. However, we identified quantitative trait loci at the *SNCA* locus, where risk alleles were associated with increased *SNCA* expression. Although evidence suggests a link between *SNCA* expression and disease risk,[23] the level of significance for this locus was not significant (p=1×10^{-4}) with the threshold for significance that we set (p<3·55×10^{-5}).

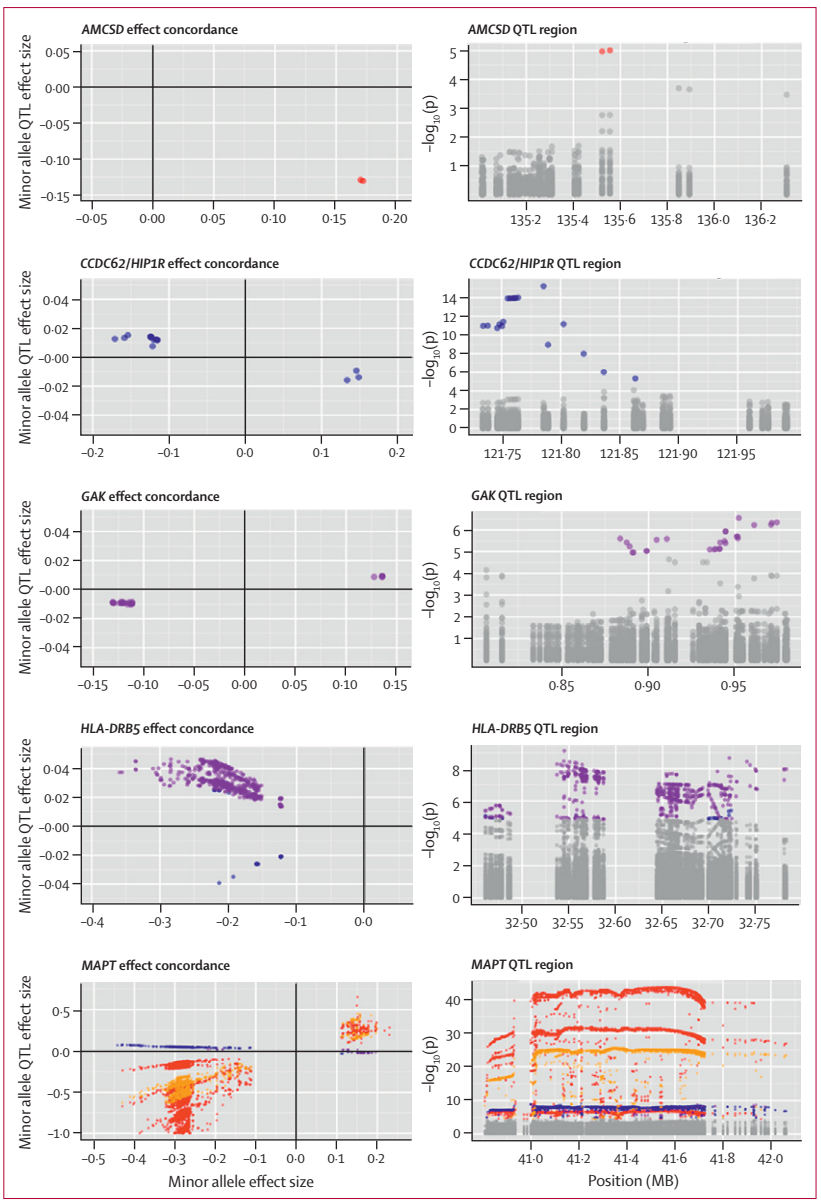


Figure 2: **Summary of quantitative trait locus analyses**
The left column shows the concordance between meta-analysis effect estimates and QTL effect estimates for SNPs at five loci with significant QTL associations. The right column shows the position of significantly associated SNPs from the QTL analyses within every region of interest. Orange circles=expression assayed in the frontal cortex. Red circles=expression assayed in the cerebellum. Purple circles=methylation assayed in the frontal cortex. Blue circles=methylation in the cerebellum. Grey circles=associations tested but that did not pass the Bonferroni correction threshold for significance (only plotted in the bottom row). QTL=quantitative trait locus. SNP=single nucleotide polymorphism.

Discussion

SNCA, *MAPT*, and *HLA-DRB5* have been confirmed as risk loci for Parkinson's disease by previous GWAS[1–7,9] and by our meta-analysis. We have also shown that, although previous GWAS were individually underpowered to prove the associations between the *BST1*, *LRRK2*, and *GAK* loci and Parkinson's disease, our meta-analysis and replication analysis identified an association at these loci (panel).[1–3,5,6,9]

GWAS investigate loci that often contain several genes and we should be mindful not to ascribe disease risk to any one gene within this locus in the absence of further biological evidence. However, the novel loci detected include biologically plausible candidate for Parkinson's

Panel: Research in context

Systematic review

As part of the International Parkinson Disease Genomics Consortium we had access to several genome-wide association datasets for Parkinson's disease, including both published and unpublished studies. To identify other studies, we searched PubMed with no date restrictions for published genome-wide association studies in Parkinson's disease with the following terms: Parkinson's disease, genome wide association. The search returned studies already included through the consortium members.[1–3,9,11] For those studies outside the existing framework of the consortium, we used all studies that included more than 300 000 single nucleotide polymorphisms and for which the underlying genetic data were available for download. This resulted in one additional dataset being added to the meta-analysis.[5]

Interpretation

Up to now, to our knowledge this study is the largest genetic analysis of Parkinson's disease undertaken and has confirmed the associations at six previously implicated loci and also identified five new loci. This study provides evidence that common genetic variation plays an important part in the cause of Parkinson's disease. We have confirmed a strong genetic component to Parkinson's disease, which, until recently, was thought to be completely caused by environmental factors. The genomic loci described show the rapid pace of growth in the specialty of genome-wide association studies of complex disease, and the future predictive use of genes identified in such studies.

disease risk. *ACMSD* is associated with picolinic and quinolinic acid homoeostasis and is a possible therapeutic target for several disorders that affect the CNS.[24] The locus identified near *STK39* has been associated with autism, hypertension, and inflammatory status,[25–27] although there have been no reports of this locus contributing to neurodegenerative phenotypes. The *LAMP3* locus might partly cause modulation of neuronal and neurosecretory function in PC12 cell lines.[28] *HLA-DRB5* is associated with multiple sclerosis, immunocompetence, and histocompatibility.[29–31] The association with Parkinson's disease at *HLA-DRB5* supports the theory that inflammatory factors are associated with the pathogenesis of Parkinson's disease.[32] The protein product of *HIP1R* is functionally involved in intrinsic cell-death pathways and interacted with huntingtin to modulate polyglutamine-induced neuronal dysfunction in transgenic worm and mouse models.[33] Finally, the association detected at the *SYT11* locus includes a gene that has been investigated previously in a negative mutation screening study in 393 patients with familial or sporadic Parkinson's disease[34] and in a cell biology study that showed an interaction between the protein products of *SYT11* and *PARK2* in patients with Parkinson's disease.[35]

The association between genetic variability at the *LRRK2* locus and Parkinson's disease is mechanistically interesting because data suggest that this association is a result of variability outside the common G2019S mutation, which raises the possibility that splicing or expression of wild-type *LRRK2* might be pathologically important. If this suggestion is correct, the role of *LRRK2* in Parkinson's disease might relate to an exaggeration of its normal function rather than some gain of abnormal function.

Understanding of the pathobiologically relevant effect of the identified risk variants in Parkinson's disease is challenging. However, we associated changes in expression and DNA methylation with risk alleles at five of the identified loci. This work has many caveats, not least of which is that it is associative and does not imply causality. However, these data do serve as a launching point for further investigation into the biological basis of Parkinson's disease.

The absence of predictive capacity in the risk-profile estimates suggests that common genetic variability at these loci, the small risk estimates per locus in this meta-analysis (and GWAS-based studies in general), and the inability to include putative functional variants per locus, do not allow clinically relevant predictive power to be quantified. Additionally, no environmental factors were included in risk profiling or population-attributable risk estimates, which might have led to some overestimation of the genetic risk of Parkinson's disease, because its cause is probably not entirely genetic.

Assumptions are unavoidable when modelling population-attributable risk with data from GWAS. Thus, we have probably overestimated the genetic component of Parkinson's disease risk on the basis of these loci alone because bias inherent in using a case-control study will slightly skew the frequency of risk alleles higher. However, this calculation did allow us to rank the contribution of every locus to the genetic cause of Parkinson's disease, and to estimate the possible decrease in the future incidence of Parkinson's disease achieved by preventative treatments targeted at genetic causes. Risk-profile modelling provides a conservative estimate of genetic risk and has moderate predictive power. The identification of additional common and rare risk variants for Parkinson's disease will probably revise our estimate of the genetic component of disease upward.

Contributors

MAN, VP, JS-S, MM, JH, PH, AB, TG, ABS, and NWW designed the study. MAN, MM, JH, PH, AB, TG, ABS, and NWW obtained funding. DGH, MSh, U-MS, JS-S, CS, SL, SS, KS, MM, JH, PH, AB, TG, ABS, and NWW collected samples. KS, MM, JH, PH, AB, TG, ABS, and NWW supervised the study. MAN, VP, MSh, U-MS, MSa, JS-S, CS, SL, MM, AB, and ABS did the data analysis and data management.

International Parkinson Disease Genomics Consortium members

Michael A Nalls (Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA), Vincent Plagnol (UCL Genetics Institute, London, UK), Dena G Hernandez (Laboratory of Neurogenetics, National Institute on

Aging; and Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK), Manu Sharma (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, University of Tübingen, and DZNE, German Center for Neurodegenerative Diseases, Tübingen, Germany), Una-Marie Sheerin (Department of Molecular Neuroscience, UCL Institute of Neurology), Mohamad Saad (INSERM U563, CPTP, Toulouse, France; and Paul Sabatier University, Toulouse, France), Javier Simón-Sánchez (Department of Clinical Genetics, Section of Medical Genomics, VU University Medical Centre, Amsterdam, Netherlands), Claudia Schulte (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research), Suzanne Lesage (INSERM, UMR_S975 [formerly UMR_S679], Paris, France; Université Pierre et Marie Curie-Paris, Centre de Recherche de l'Institut du Cerveau et de la Moelle épinière, Paris, France; and CNRS, Paris, France), Sigurlaug Sveinbjörnsdóttir (Department of Neurology, Landspítali University Hospital, Reykjavík, Iceland; Department of Neurology, MEHT Broomfield Hospital, Chelmsford, Essex, UK; and Queen Mary College, University of London, London, UK), Sampath Arepalli (Laboratory of Neurogenetics, National Institute on Aging), Roger Barker (Department of Neurology, Addenbrooke's Hospital, University of Cambridge, Cambridge, UK), Yoav Ben-Shlomo (Department of Social Medicine, Bristol University, UK), Henk W Berendse (Department of Neurology and Alzheimer Center, VU University Medical Center), Daniela Berg (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research), Kailash Bhatia (Department of Motor Neuroscience, UCL Institute of Neurology), Rob M A de Bie (Department of Neurology, Academic Medical Center, University of Amsterdam, Amsterdam, Netherlands), Alessandro Biffi (Center for Human Genetic Research and Department of Neurology, Massachusetts General Hospital, Boston, MA, USA; and Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA), Bas Bloem (Department of Neurology, Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands), Zoltan Bochdanovits (Department of Clinical Genetics, Section of Medical Genomics, VU University Medical Centre), Michael Bonin (Department of Medical Genetics, Institute of Human Genetics, University of Tübingen, Tübingen, Germany), Jose M Bras (Department of Molecular Neuroscience, UCL Institute of Neurology), Kathrin Brockmann (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research), Janet Brooks (Laboratory of Neurogenetics, National Institute on Aging), David J Burn (Newcastle University Clinical Ageing Research Unit, Campus for Ageing and Vitality, Newcastle upon Tyne, UK), Gavin Charlesworth (Department of Molecular Neuroscience, UCL Institute of Neurology), Honglei Chen (Epidemiology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, NC, USA), Patrick F Chinnery (Neurology M4104, The Medical School, Framlington Place, Newcastle upon Tyne, UK), Sean Chong (Laboratory of Neurogenetics, National Institute on Aging), Carl E Clarke (School of Clinical and Experimental Medicine, University of Birmingham, Birmingham, UK; and Department of Neurology, City Hospital, Sandwell and West Birmingham Hospitals NHS Trust, Birmingham, UK), Mark R Cookson (Laboratory of Neurogenetics, National Institute on Aging), J Mark Cooper (Department of Clinical Neurosciences, UCL Institute of Neurology), Jean Christophe Corvol (INSERM, UMR_S975; Université Pierre et Marie Curie-Paris; CNRS; and INSERM CIC-9503, Hôpital Pitié-Salpêtrière, Paris, France), Carl Counsell (University of Aberdeen, Division of Applied Health Sciences, Population Health Section, Aberdeen, UK), Philippe Damier (CHU Nantes, CIC0004, Service de Neurologie, Nantes, France), Jean-François Dartigues (INSERM U897, Université Victor Segalen, Bordeaux, France), Panos Deloukas (Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, UK), Günther Deuschl (Klinik für Neurologie, Universitätsklinikum Schleswig-Holstein, Campus Kiel, Christian-Albrechts-Universität Kiel, Kiel, Germany), David T Dexter (Parkinson's Disease Research Group, Faculty of Medicine, Imperial College London, London, UK), Karin D van Dijk (Department of Neurology and Alzheimer Center, VU University Medical Center), Allissa Dillman (Laboratory of Neurogenetics, National Institute on Aging), Frank Durif (Service de Neurologie, Hôpital Gabriel Montpied, Clermont-Ferrand, France), Alexandra Dürr (INSERM, UMR_S975; Université Pierre et Marie Curie-Paris; CNRS; and AP-HP, Pitié-Salpêtrière Hospital), Sarah Edkins (Wellcome Trust Sanger Institute), Jonathan R Evans (Cambridge Centre for Brain Repair, Cambridge, UK), Thomas Foltynie (UCL Institute of Neurology), Jianjun Gao (Epidemiology Branch, National Institute of Environmental Health Sciences), Michelle Gardner (Department of Molecular Neuroscience, UCL Institute of Neurology), J Raphael Gibbs (Laboratory of Neurogenetics, National Institute on Aging; and Department of Molecular Neuroscience, UCL Institute of Neurology), Alison Goate (Department of Psychiatry, Department of Neurology, Washington University School of Medicine, MI, USA), Emma Gray (Wellcome Trust Sanger Institute), Rita Guerreiro (Department of Molecular Neuroscience, UCL Institute of Neurology), Ómar Gústafsson (deCODE genetics and Department of Psychiatry, Oslo University Hospital, N-0407 Oslo, Norway), Clare Harris (University of Aberdeen), Jacobus J van Hilten (Department of Neurology, Leiden University Medical Center, Leiden, Netherlands), Albert Hofman (Department of Epidemiology, Erasmus University Medical Center, Rotterdam, Netherlands), Albert Hollenbeck (AARP, Washington DC, USA), Janice Holton (Queen Square Brain Bank for Neurological Disorders, UCL Institute of Neurology), Michele Hu (Department of Clinical Neurology, John Radcliffe Hospital, Oxford, UK), Xuemei Huang (Departments of Neurology, Radiology, Neurosurgery, Pharmacology, Kinesiology, and Bioengineering, Pennsylvania State University–Milton S Hershey Medical Center, Hershey, PA, USA), Heiko Huber (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research), Gavin Hudson (Neurology M4104, The Medical School, Newcastle upon Tyne, UK), Sarah E Hunt (Wellcome Trust Sanger Institute), Johanna Huttenlocher (deCODE genetics), Thomas Illig (Institute of Epidemiology, Helmholtz Zentrum München, German Research Centre for Environmental Health, Neuherberg, Germany), Pálmi V Jónsson (Department of Geriatrics, Landspítali University Hospital, Reykjavík, Iceland), Jean-Charles Lambert (INSERM U744, Lille, France; and Institut Pasteur de Lille, Université de Lille Nord, Lille, France), Cordelia Langford (Cambridge Centre for Brain Repair), Andrew Lees (Queen Square Brain Bank for Neurological Disorders), Peter Lichtner (Institute of Human Genetics, Helmholtz Zentrum München, German Research Centre for Environmental Health, Neuherberg, Germany), Patricia Limousin (Institute of Neurology, Sobell Department, Unit of Functional Neurosurgery, London, UK), Grisel Lopez (Section on Molecular Neurogenetics, Medical Genetics Branch, NHGRI, National Institutes of Health), Delia Lorenz (Klinik für Neurologie, Universitätsklinikum Schleswig-Holstein), Alisdair McNeill (Department of Clinical Neurosciences, UCL Institute of Neurology), Catriona Moorby (School of Clinical and Experimental Medicine, University of Birmingham), Matthew Moore (Laboratory of Neurogenetics, National Institute on Aging), Huw R Morris (MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University School of Medicine, Cardiff, UK), Karen E Morrison (School of Clinical and Experimental Medicine, University of Birmingham; and Neurosciences Department, Queen Elizabeth Hospital, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK), Ese Mudanohwo (Neurogenetics Unit, UCL Institute of Neurology and National Hospital for Neurology and Neurosurgery), Sean S O'Sullivan (Queen Square Brain Bank for Neurological Disorders), Justin Pearson (MRC Centre for Neuropsychiatric Genetics and Genomics), Joel S Perlmutter (Department of Neurology, Radiology, and Neurobiology at Washington University, St Louis), Hjörvar Pétursson (deCODE genetics; and Department of Medical Genetics, Institute of Human Genetics, University of Tübingen), Pierre Pollak (Service de Neurologie, CHU de Grenoble, Grenoble, France), Bart Post (Department of Neurology, Radboud University Nijmegen Medical Centre), Simon Potter (Wellcome Trust Sanger Institute), Bernard Ravina (Translational Neurology, Biogen Idec, MA, USA), Tamas Revesz (Queen Square Brain Bank for Neurological Disorders), Olaf Riess (Department of Medical Genetics, Institute of Human Genetics, University of Tübingen), Fernando Rivadeneira (Departments of Epidemiology and Internal Medicine, Erasmus University Medical Center), Patrizia Rizzu (Department of Clinical Genetics, Section of Medical Genomics, VU University Medical Centre), Mina Ryten (Department of Molecular Neuroscience, UCL Institute of Neurology), Stephen Sawcer (University of Cambridge, Department of Clinical Neurosciences, Addenbrooke's

hospital, Cambridge, UK), Anthony Schapira (Department of Clinical Neurosciences, UCL Institute of Neurology), Hans Scheffer (Department of Human Genetics, Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands), Karen Shaw (Queen Square Brain Bank for Neurological Disorders), Ira Shoulson (Department of Neurology, University of Rochester, Rochester, NY, USA), Ellen Sidransky (Section on Molecular Neurogenetics, Medical Genetics Branch, NHGRI), Colin Smith (Department of Pathology, University of Edinburgh, Edinburgh, UK), Chris C A Spencer (Wellcome Trust Centre for Human Genetics, Oxford, UK), Hreinn Stefánsson (deCODE genetics), Stacy Steinberg (deCODE genetics), Joanna D Stockton (School of Clinical and Experimental Medicine), Amy Strange (Wellcome Trust Centre for Human Genetics), Kevin Talbot (University of Oxford, Department of Clinical Neurology, John Radcliffe Hospital, Oxford, UK), Carlie M Tanner (Clinical Research Department, The Parkinson's Institute and Clinical Center, Sunnyvale, CA, USA), Avazeh Tashakkori-Ghanbaria (Wellcome Trust Sanger Institute), François Tison (Service de Neurologie, Hôpital Haut-Lévêque, Pessac, France), Daniah Trabzuni (Department of Molecular Neuroscience, UCL Institute of Neurology), Bryan J Traynor (Laboratory of Neurogenetics, National Institute on Aging), André G Uitterlinden (Departments of Epidemiology and Internal Medicine, Erasmus University Medical Center), Daan Velseboer (Department of Neurology, Academic Medical Center), Marie Vidailhet (INSERM, UMR_S975, Université Pierre et Marie Curie-Paris, CNRS, UMR 7225), Robert Walker (Department of Pathology, University of Edinburgh), Bart van de Warrenburg (Department of Neurology, Radboud University Nijmegen Medical Centre), Mirdhu Wickremaratchi (Department of Neurology, Cardiff University, Cardiff, UK), Nigel Williams (MRC Centre for Neuropsychiatric Genetics and Genomics), Caroline H Williams-Gray (Department of Neurology, Addenbrooke's Hospital), Sophie Winder-Rhodes (Department of Psychiatry and Medical Research Council and Wellcome Trust Behavioural and Clinical Neurosciences Institute, University of Cambridge), Kári Stefánsson (deCODE genetics), Maria Martinez (INSERM U563; and Paul Sabatier University), John Hardy (Department of Molecular Neuroscience, UCL Institute of Neurology), Peter Heutink (Department of Clinical Genetics, Section of Medical Genomics, VU University Medical Centre), Alexis Brice (INSERM, UMR_S975, Université Pierre et Marie Curie-Paris, CNRS, UMR 7225, AP-HP, Pitié-Salpêtrière Hospital), Wellcome Trust Case-Control Consortium 2 (webappendix p 13), Thomas Gasser (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, and DZNE, German Center for Neurodegenerative Diseases), Andrew B Singleton (Laboratory of Neurogenetics, National Institute on Aging), Nicholas W Wood (UCL Genetics Institute; and Department of Molecular Neuroscience, UCL Institute of Neurology).

Writing group

Michael A Nalls, Vincent Plagnol, Dena G Hernandez, Manu Sharma, Una-Marie Sheerin, Mohamad Saad, J Simón-Sánchez, Claudia Schulte, Suzanne Lesage, Sigurlaug Sveinbjörnsdóttir, Kári Stefánsson, Maria Martinez, John Hardy, Peter Heutink, Alexis Brice, Thomas Gasser, Andrew B Singleton, Nicholas W Wood.

Conflicts of interest

KSt has received grants from deCODE. JH has received consulting fees or honoraria from Eisai and his institute has received consulting fees or honoraria from Merck-Serono. TG has received consultancy fees from Cephalon and Merck-Serono; grants from Novartis; payments for lectures including service on speakers' bureaus from Boehringer Ingelheim, Merck-Serono, UCB, and Valeant; and holds patents NGFN2 and KASPP. MAN, VP, DGH, MSh, U-MS, MSa, JS-S, CS, SL, SS, KS, MM, PH, ABr, ABS, and NWW declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Intramural Research Programs of the National Institute on Aging, National Institute of Neurological Disorders and Stroke, National Institute of Environmental Health Sciences, and National Human Genome Research Institute of the National Institutes of Health, Department of Health and Human Services (project numbers Z01-AG000949-02 and Z01-ES101986). This work was also supported by the US Department of Defense (award number W81XWH-09-2-0128); National Institutes of Health (grants NS057105 and RR024992); American Parkinson disease Association (APDA); Barnes Jewish Hospital Foundation; Greater St Louis Chapter of the APDA; Hersenstichting Nederland; Neuroscience Campus Amsterdam; and the section of medical genomics, the Prinses Beatrix Fonds. The KORA (Cooperative Research in the Region of Augsburg) research platform was started and financed by the Forschungszentrum für Umwelt und Gesundheit, which is funded by the German Federal Ministry of Education, Science, Research, and Technology and by the State of Bavaria. This study was also funded by the German National Genome Network (NGFNplus number 01GS08134, German Ministry for Education and Research); by the German Federal Ministry of Education and Research (NGFN 01GR0468, PopGen); and 01EW0908 in the frame of ERA-NET NEURON and Helmholtz Alliance Mental Health in an Ageing Society (HA-215), which was funded by the Initiative and Networking Fund of the Helmholtz Association. The French GWAS work was supported by the French National Agency of Research (ANR-08-MNP-012). This study was also sponsored by the Landspitali University Hospital Research Fund (grant to SSv); Icelandic Research Council (grant to SSv); and European Community Framework Programme 7, People Programme, and IAPP on novel genetic and phenotypic markers of Parkinson's disease and Essential Tremor (MarkMD), contract number PIAP-GA-2008-230596 MarkMD (to HP and JHu). We used the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD, USA, and DNA panels, samples, and clinical data from the National Institute of Neurological Disorders and Stroke Human Genetics Resource Center DNA and Cell Line Repository. People who contributed samples are acknowledged in descriptions of every panel on the repository website. We thank The French Parkinson's Disease Genetics Study Group: Y Agid, M Anheim, A-M Bonnet, M Borg, A Brice, E Broussolle, J-C Corvol, P Damier, A Destée, A Dürr, F Durif, S Klebe, E Lohmann, M Martinez, P Pollak, O Rascol, F Tison, C Tranchant, M Vérin, F Viallet, and M Vidailhet. We also thank the members of the French 3C Consortium: Annick Alpérovitch, Claudine Berr, Christophe Tzourio, and Philippe Amouyel for allowing us to use part of the 3C cohort; and D Zelenika for support in generating the genome-wide molecular data. We used genome-wide association data generated by the Wellcome Trust Case-Control Consortium 2 (WTCCC2) from UK patients with Parkinson's disease and UK control individuals from the 1958 Birth Cohort and National Blood Service. Genotyping of UK replication cases on ImmunoChip was part of the WTCCC2 project, which was funded by the Wellcome Trust (083948/Z/07/Z). UK population control data was made available through WTCCC1. This study was supported by the Medical Research Council and Wellcome Trust disease centre (grant WT089698/Z/09/Z to NW, JHa, and ASc). This study was also supported by Parkinson's UK (grants 8047 and J-0804) and the Medical Research Council (G0700943). We thank Jeffrey Barrett for assistance with the design of the ImmunoChip. DNA extraction work that was done in the UK was undertaken at University College London Hospitals, University College London, who received a proportion of funding from the Department of Health's National Institute for Health Research Biomedical Research Centres funding. This study was supported in part by the Wellcome Trust/Medical Research Council Joint Call in Neurodegeneration award (WT089698) to the Parkinson's Disease Consortium (UKPDC), whose members are from the UCL Institute of Neurology, University of Sheffield, and the Medical Research Council Protein Phosphorylation Unit at the University of Dundee.

References

1 Saad M, Lesage S, Saint-Pierre A, et al, for the French Parkinson's Disease Genetics Study Group. Genome-wide association study confirms BST1 and suggests a locus on 12q24 as risk loci for Parkinson's disease in the European population. *Hum Mol Genet* 2011; **20:** 615–27.

2 Simon-Sanchez J, Schulte C, Bras JM, et al. Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat Genet* 2009; **41:** 1308–12.

3 The UK Parkinson's Disease Consortium and The Wellcome Trust Case Control Consortium 2. Dissection of the genetics of Parkinson's disease identifies an additional association 5´ of SNCA and multiple associated haplotypes at 17q21. *Hum Mol Genet* 2011; **20:** 345–53.

4 Edwards TL, Scott WK, Almonte C, et al. Genome-wide association study confirms SNPs in SNCA and the MAPT region as common risk factors for Parkinson disease. *Ann Hum Genet* 2010; **74:** 97–109.

5 Pankratz N, Wilk JB, Latourelle JC, et al. Genomewide association study for susceptibility genes contributing to familial Parkinson disease. *Hum Genet* 2009; **124:** 593–605.

6 Satake W, Nakabayashi Y, Mizuta I, et al. Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. *Nat Genet* 2009; **41:** 1303–07.

7 Hamza TH, Zabetian CP, Tenesa A, et al. Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. *Nat Genet* 2010; **42:** 781–85.

8 Singleton AB, Hardy J, Traynor BJ, Houlden H. Towards a complete resolution of the genetic architecture of disease. *Trends Genet* 2010; **26:** 438–42.

9 Simón-Sánchez J, van Hilten JJ, van de Warrenburg B, et al. Genome-wide association study confirms extant PD risk loci among the Dutch. *Eur J Hum Genet* 2011; published online Jan 19. DOI:10.1038/ejhg.2010.254

10 Mailman MD, Feolo M, Jin Y, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* 2007; **39:** 1181–86.

11 Fung HC, Scholz S, Matarin M, et al. Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol* 2006; **5:** 911–16.

12 Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. *Annu Rev Genomics Hum Genet* 2009; **10:** 387–406.

13 University of Michigan Center for Statistical Genetics. 1000G 2009–08 download. http://www.sph.umich.edu/csg/abecasis/MACH/download/1000G-Sanger-0908.html (accessed Dec 1, 2009).

14 Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003; **327:** 557–60.

15 Ioannidis JP, Patsopoulos NA, Evangelou E. Heterogeneity in meta-analyses of genome-wide association investigations. *PLoS One* 2007; **2:** e841.

16 Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010; **26:** 2190–91.

17 University of Michigan Centre for Statistical Genetics. Metal—meta-analysis helper. http://www.sph.umich.edu/csg/abecasis/metal (accessed Dec 1, 2009).

18 de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* 2008; **17:** R122–28.

19 R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2008.

20 Zollner S, Pritchard JK. Overcoming the winner's curse: estimating penetrance parameters from case-control data. *Am J Hum Genet* 2007; **80:** 605–15.

21 Gibbs JR, van der Brug MP, Hernandez DG, et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet* 2010; **6:** e1000952.

22 Sidransky E, Nalls MA, Aasly JO, et al. Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. *N Engl J Med* 2009; **361:** 1651–61.

23 Singleton A, Gwinn-Hardy K. Parkinson's disease and dementia with Lewy bodies: a difference in dose? *Lancet* 2004; **364:** 1105–07.

24 Garavaglia S, Perozzi S, Galeazzi L, Raffaelli N, Rizzi M. The crystal structure of human alpha-amino-beta-carboxymuconate-epsilon-semialdehyde decarboxylase in complex with 1,3-dihydroxyacetonephosphate suggests a regulatory link between NAD synthesis and glycolysis. *FEBS J* 2009; **276:** 6615–23.

25 Ramoz N, Cai G, Reichert JG, Silverman JM, Buxbaum JD. An analysis of candidate autism loci on chromosome 2q24-q33: evidence for association to the STK39 gene. *Am J Med Genet B Neuropsychiatr Genet* 2008; **147B:** 1152–58.

26 Wang Y, O'Connell JR, McArdle PF, et al. From the cover: whole-genome association study identifies STK39 as a hypertension susceptibility gene. *Proc Natl Acad Sci USA* 2009; **106:** 226–31.

27 Cunnington MS, Kay C, Avery PJ, Mayosi BM, Koref MS, Keavney B. STK39 polymorphisms and blood pressure: an association study in British caucasians and assessment of cis-acting influences on gene expression. *BMC Med Genet* 2009; **10:** 135.

28 Malosio ML, Giordano T, Laslop A, Meldolesi J. Dense-core granules: a specific hallmark of the neuronal/neurosecretory cell phenotype. *J Cell Sci* 2004; **117:** 743–49.

29 Lincoln MR, Ramagopalan SV, Chao MJ, et al. Epistasis among HLA-DRB1, HLA-DQA1, and HLA-DQB1 loci determines multiple sclerosis susceptibility. *Proc Natl Acad Sci USA* 2009; **106:** 7542–47.

30 Cano P, Klitz W, Mack SJ, et al. Common and well-documented HLA alleles: report of the ad-hoc committee of the American Society for Histocompatiblity and Immunogenetics. *Hum Immunol* 2007; **68:** 392–417.

31 Handel AE, Handunnetthi L, Berlanga AJ, Watson CT, Morahan JM, Ramagopalan SV. The effect of single nucleotide polymorphisms from genome wide association studies in multiple sclerosis on gene expression. *PLoS One* 2010; **5:** e10142.

32 Wersinger C, Sidhu A. An inflammatory pathomechanism for Parkinson's disease? *Curr Med Chem* 2006; **13:** 591–602.

33 Parker JA, Metzler M, Georgiou J, et al. Huntingtin-interacting protein 1 influences worm and mouse presynaptic function and protects *Caenorhabditis elegans* neurons against mutant polyglutamine toxicity. *J Neurosci* 2007; **27:** 11056–64.

34 Glass AS, Huynh DP, Franck T, et al. Screening for mutations in synaptotagmin XI in Parkinson's disease. *J Neural Transm Suppl* 2004; **68:** 21–28.

35 Huynh DP, Scoles DR, Nguyen D, Pulst SM. The autosomal recessive juvenile Parkinson disease gene product, parkin, interacts with and ubiquitinates synaptotagmin XI. *Hum Mol Genet* 2003; **12:** 2587–97.

OPEN ACCESS Freely available online

PLoS GENETICS

A Two-Stage Meta-Analysis Identifies Several New Loci for Parkinson's Disease

International Parkinson's Disease Genomics Consortium (IPDGC), Wellcome Trust Case Control Consortium 2 (WTCCC2)¶

Abstract

A previous genome-wide association (GWA) meta-analysis of 12,386 PD cases and 21,026 controls conducted by the International Parkinson's Disease Genomics Consortium (IPDGC) discovered or confirmed 11 Parkinson's disease (PD) loci. This first analysis of the two-stage IPDGC study focused on the set of loci that passed genome-wide significance in the first stage GWA scan. However, the second stage genotyping array, the ImmunoChip, included a larger set of 1,920 SNPs selected on the basis of the GWA analysis. Here, we analyzed this set of 1,920 SNPs, and we identified five additional PD risk loci (combined $p<5\times10^{-10}$, *PARK16*/1q32, *STX1B*/16p11, *FGF20*/8p22, *STBD1*/4q21, and *GPNMB*/7p15). Two of these five loci have been suggested by previous association studies (*PARK16*/1q32, *FGF20*/8p22), and this study provides further support for these findings. Using a dataset of post-mortem brain samples assayed for gene expression (n = 399) and methylation (n = 292), we identified methylation and expression changes associated with PD risk variants in *PARK16*/1q32, *GPNMB*/7p15, and *STX1B*/16p11 loci, hence suggesting potential molecular mechanisms and candidate genes at these risk loci.

Citation: International Parkinson's Disease Genomics Consortium (IPDGC), Wellcome Trust Case Control Consortium 2 (WTCCC2) (2011) A Two-Stage Meta-Analysis Identifies Several New Loci for Parkinson's Disease. PLoS Genet 7(6): e1002142. doi:10.1371/journal.pgen.1002142

Editor: Greg Gibson, Georgia Institute of Technology, United States of America

Received March 1, 2011; **Accepted** April 8, 2011; **Published** June 30, 2011

Funding: This work was supported in part by the Wellcome Trust/MRC Joint Call in Neurodegeneration award (WT089698) to the UK Parkinson's Disease Consortium (UKPDC), whose members are from the UCL/Institute of Neurology, the University of Sheffield, and the MRC Protein Phosphorylation Unit at the University of Dundee. Additionally, part of the study was undertaken at UCLH/UCL using funding through a Department of Health NIHR Biomedical Research Centre. This work was also supported by Parkinson's UK (Grants 8047 and J-0804) and the Medical Research Council (G0700943). Genotyping of UK replication cases on Immunochip was part of the Wellcome Trust Case Control Consortium 2 project which is funded by the Wellcome Trust (085475/B/08/Z and 085475/Z/08/Z). P Damier is partly supported by a Wolfson-Royal Society Merit award. The UK gene expression work was supported in part by the UK Medical Research Council (G0901254) to researchers based in the UCL Institute of Neurology and King's College London. J Holton receives support from the Reta Lila Weston Trust for Medical Research. This work was also supported by the Landspitali University Hospital Research Fund (S Sveinbjörnsdóttir), the Icelandic Research Council (S Sveinbjörnsdóttir), the European Community Framework Programme 7, People programme, IAPP on novel genetic and phenotypic markers of Parkinson's disease, and Essential Tremor (MarkMD), contract no PIAP-GA-2008-230596 MarkMD (H Pétursson, J Holton). This US work was supported in part by the Intramural Research Programs of the National Institute on Aging, National Institute of Neurological Disorders and Stroke, National Institute of Environmental Health Sciences, National Human Genome Research Institute, National Institutes of Health, Department of Health and Human Services; project numbers Z01 AG000949-02 and Z01-ES101986. In addition this study was supported by the US Department of Defense, award number W81XWH-09-2-0128. Funding to support collection of a portion of the samples was obtained from the National Institutes of Health (grants NS057105 and RR024992), the American Parkinson Disease Association (APDA), Barnes Jewish Hospital Foundation, and the Greater St. Louis Chapter of the APDA. The KORA research platform (KORA: Cooperative Research in the Region of Augsburg; http://www.gsf.de/KORA) was initiated and financed by the ForschungszentrumfürUmwelt und Gesundheit (GSF), which is funded by the German Federal Ministry of Education, Science, Research, and Technology and by the State of Bavaria. The study was additionally funded by the German National Genome Network (NGFNplus #01GS08134; German Ministry for Education and Research) and by the German Federal Ministry of Education and Research (BMBF) NGFN (01GR0468) and in the frame of ERA-Net NEURON (01GW0908). This work was also supported by the Helmholtz Alliance Mental Health in an Ageing Society (HelMA, HA-215) funded by the Initiative and Networking Fund of the Helmholtz Association. The French GWA scan work was supported by the French National Agency of Research (http://www.agence-nationale-recherche.fr, ANR-08-MNP-012) and by the National Research Funding Agency (ANR-08-NEUR-004-01) in ERA-NET NEURON framework (http://www.neuron-eranet.eu). We also want to thank the Hersenstichting Nederland (http://www.hersenstichting.nl), the Neuroscience Campus Amsterdam and the section of Medical genomics, the Prinses Beatrix Fonds (http://www.prinsesbeatrixfonds.nl) for sponsoring this work. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: n.wood@ion.ucl.ac.uk

¶ A full list of authors and affiliations is provided in the Acknowledgments. A full list of members of the Wellcome Trust Case Control Consortium 2 (WTCCC2) consortium is provided in Text S1.

Introduction

Until the recent developments of high throughput genotyping and genome-wide association (GWA) studies, little was known of the genetics of typical Parkinson's disease (PD). Studies of the genetic basis of familial forms of PD first identified rare highly penetrant mutations in *LRKK2* [1,2], *PINK1* [3], *SNCA* [4], *PARK2* [5] and *PARK7* [6]. Following these findings, GWA scans for idiopathic PD identified *SNCA* and *MAPT* as unequivocal risk loci [7,8,9,10,11] as well as implicated *BST1* [8], *GAK* [12], and *HLA-DR* [13]. Using sequence based imputation methods [14], the meta-analysis of several GWA scans [7,9,10,11] conducted by the International Parkinson's Disease Genomics Consortium (IPDGC) identified and replicated five new loci: *ACMSD*, *STK39*, *MCCC1/LAMP3*, *SYT11*, and *CCDC62/HIP1R* [15] and confirmed association at *SNCA*, *LRRK2*, *MAPT*, *BST1*, *GAK* and *HLA-DR* [15].

We conducted a two-stage association study. Combining stage 1 and stage 2, the data consist of 12,386 PD cases and 21,026 controls genotyped using a variety of platforms (Table 1). Stage 1 used genome-wide genotyping arrays and our initial analysis [15] focused on the subset of SNPs that passed genome-wide significance in stage 1. For stage 2 genotyping, we used a custom content Illumina iSelect array, the ImmunoChip and additional

Author Summary

This paper describes the largest case-control analysis of Parkinson's disease to date, with a combined sample set of over 12,000 cases and 21,000 controls. After combining our findings with an independent replication dataset of more than 3,000 cases and 29,000 controls, we found five additional PD risk loci in addition to the 11 loci previously identified in earlier consortium efforts. This successful study further demonstrates the power of the GWA scan experimental design to find new loci contributing to disease risk, even in the context of complex disorders like Parkinson's disease. These new findings provide insights into the etiology of PD and will promote a better understanding of its pathogenesis.

GWAS typing as previously described [15]. The primary content of the ImmunoChip data focuses on autoimmune disorders but, as part of a collaborative agreement with the Wellcome Trust Case Control Consortium 2, we included 1,920 ImmunoChip SNPs on the basis of the stage 1 GWA PD results.

Here, we report the combined analysis for this full set of 1,920 SNPs. This step1+2 analysis identified seven new loci that passed genome-wide significance in the meta-analysis. During the process of analyzing these data and preparing for publication, we became aware that another group was also preparing a large independent GWA scan in PD for publication (Do et al, submitted). Following discussion with this group we agreed to cross validate the top hits from each study by exchanging summary statistics for this small number of loci.

To provide further insights into the molecular function of these associated variants, we tested risk alleles at these loci for correlation with the expression of physically close gene (expression quantitative trait locus, eQTL) and the methylation status (methQTL) of proximal DNA CpG sites in a dataset of 399 control frontal cortex and cerebellar tissue samples extracted post-mortem from individuals without a history of neurological disorders.

Results

In addition to eleven loci that passed genome-wide significance in stage 1 [15], we identified over 100 regions of interest defined as 10 kb windows containing at least one SNP associated at $p<10^{-3}$. We submitted the most associated SNP in each region for probe design and follow-up genotyping using the ImmunoChip platform. For each region of interest, we also added four SNPs in high level of linkage disequilibrium (LD) to provide redundancy where the most associated SNP would not pass the Illumina probe design step or the assay for that SNP would fail. To complete the array design we also added all non-synonymous dbSNPs located in known PD associated regions [1,2,3,4,5,6]. Out of these 2,400 submitted SNPs, 1,920 passed QC and were included in the final array design. For these 1,920 SNPs we combined stage 1 and stage 2 associated data in a meta-analysis of 12,386 cases and 21,026 controls (Table 1) from the IPDGC. We exchanged summary statistics for these most significant hits with an additional large, case-control replication dataset (3,426 PD cases and 29,624 controls) in an attempt to demonstrate independent replication.

On the basis of stage 1+2 results, seven new SNPs passed our defined genome-wide significance threshold ($p<5\times10^{-8}$, Table 2 and Figure 1). These loci are either novel or the previous evidence of association was not entirely convincing in individuals of European descent. We combined these results with the independent replication. Five of these seven loci replicated and showed strong combined evidence of PD association ($p<10^{-10}$ overall). Taking either the nearest gene (or the strongest candidate when available) to designate these regions, these five loci are 1q32/*PARK16* [7], 4q21/*STBD1*, 7p15/*GPNMB*, 8p22/*FGF20* [16] and 16p11/*STX1B*.

rs708723/1q32 has been previously reported as PD associated (*PARK16*, [7,8]) but this SNP lacked the unequivocal evidence of association in European samples ($p=9.47\times10^{-10}$ in stage 2 only). To understand the potential biological consequences of risk variation at this locus we tested whether rs708723 was correlated with either gene expression or DNA methylation status of proximal transcripts or CpG sites respectively (Table 3). We found correlations with the expression of *NUCKS1* ($p=1.8\times10^{-7}$) and *RAB7L1* ($p=7.2\times10^{-4}$). We also found correlations with the methylation state of CpG sites located in the *FLJ3269* gene ($p=3.9\times10^{-22}$).

In the case of 16p11/*STX1B*, the proximal gene to the most associated SNP rs4889603 is *SETD1A*. However, *STX1B* is located 18 kb upstream of rs4889603 and is a more plausible PD candidate gene [17] owing to its synaptic receptor function. We therefore used this gene to designate this region. Our methQTL/eQTL dataset identified a correlation between the rs4889603 risk allele and increased methylation of a CpG dinucleotide in *STX1B* (Table 3).

The SNP rs591323 in the 8p22 region is located ~150 kb downstream of the *FGF20* gene (NCBI build 36.3), for which association with PD has been suggested previously in familial PD samples [16,18] but which remained controversial [19]. Our findings provide further support for a PD association at this locus, but again, whether the functionally affected transcript is *FGF20* or not remains unclear.

The regions 4q21/*STBD1* and 7p15/*GPNMD* have not been previously implicated in PD etiology. We found that the risk allele of rs156429, the most associated SNP in the 7p15 region, is associated in our eQTL dataset with decreased expression of the proximal transcript encoded by *NUPL2* (Table 3). The same risk allele is also associated with increased methylation of multiple CpG sites proximal to *GPNMB* itself (Table 3). Neither of these regions contains an obvious candidate gene.

Two additional loci (3q26/*NMD3* and 8q21/*MMP16*) showed strong evidence of association in stage 1 and 2 but were not disease associated in the Do et al dataset. Further replication is required to clarify the role of variation at these loci in risk for PD.

The strongly associated G2019S variant in the *LRRK2* gene [20] was included in the Immunochip design and we replicated the published association: control frequency: 0.045% case frequency 0.61%, estimated odds ratio: 13.5 with 95% confidence interval: 5.5–43. However, the case collections have been partially screened for this variant therefore its frequency in cases and the odds ratio is likely to be underestimated.

The ImmunoChip array design provides some power to detect whether multiple distinct association signals exist at individual loci. Indeed, if a SNP showed an independent and sufficiently strong association in stage 1, it would have been included in stage 2 provided that it was not located in the same 10 kb window as the primary SNP in the region. There is precedent for this in PD, with the previous identification of independent risk signals at the *SNCA* locus [11]. We therefore used the Immunochip data to test whether any of the seven loci in Table 2 showed some evidence of more than one independent signal. None of these seven loci showed any association ($p>0.01$) after conditioning on the main SNP in the region. In contrast, after conditioning on the most associated SNPs rs356182 in the *SNCA* region, several SNPs

remained convincingly associated ($p = 9.7 \times 10^{-8}$ for rs2245801 being the most significant).

Lastly, we performed a risk profile analysis to investigate the power to discriminate cases and controls on the basis of the 16 confirmed common associated variants (Table 4). For each locus, we estimated the odds ratio on the basis of stage 1 data and we applied these estimates to compute for each individual in the ImmunoChip cohort a combined risk score. Solely based on these 16 common variants, and therefore not considering rare highly penetrant variants such as G2019S in *LRKK2* [20], we found that individuals in the top quintile of the risk score have an estimated three-fold increase in PD risk compared to individuals in the bottom quintile (Table 4). We note however that the effect size of several of these associated variants could be over-estimated (an effect known as winner's curse, see [21]) but given the consistent estimates of odds ratio across studies (Table 4) we expect this bias to be minimal.

Discussion

The combination of GWA scans and imputation methods in large cohorts of PD cases and controls has enabled us to identify five PD associated loci in addition to the 11 previously reported by us. Two of these loci (1q32/*PARK16*, 8p22/*FGF20*) implicate regions that had been previously associated with PD risk [8,16]. The 1q32/*PARK16* showed convincing evidence of association in the Japanese population [8] but until now the association P-value had not passed a stringent genome-wide significance threshold in samples of European descent [7]. The 8p22/*FGF20* locus had been previously reported in a study of familial PD [16] and we provide the first evidence of association in a case-control study. The remaining three loci (*STX1B*/16p11, *STBD1*/4q21 and *GPNMB*/7p15) are new.

Adding the eleven previously reported common variants [15] to the five convincingly associated loci identified in this study, common variants at 16 loci have now been associated with PD. Controlling for the risk score based on the 11 SNPs previously identified [15] in the risk profile analysis (Table 4), the addition of these five new loci provides a modest but significant ($p = 2.2 \times 10^{-3}$) improvement of our ability to discriminate PD cases from controls.

Combining eQTL/methylation and case-control data implicates potential mechanisms which could explain the increased PD risk associated some of these variants. In particular, the strong eQTL in the 1q32/*PARK16* region with the *RAB7L1* and *NUCKS1* genes (Table 3) suggests that either one of these genes could be the biological effector of this risk locus. However, existing data show that eQTLs are widespread and this co-localization could be the result of chance alone [22]. Additional fine-mapping work will be required to assess whether the expression and case-control data are indeed fully consistent.

While we are unable to unequivocally pinpoint the causative genes underlying these associations, their known biological function can suggest likely candidates. At the 1q32/*PARK16* loci our association and eQTL data indicate that *RAB7L1* and *NUCKS1* are the best candidates. The former is a GTP-binding protein that plays an important role in the regulation of exocytotic and endocytotic pathways [23]. Exocytosis is relevant for PD for two main reasons: firstly, since dopaminergic neurotransmission is mediated by the vesicular release of dopamine, i.e. dopamine exocytosis [24], and secondly because it has been shown that alpha-synuclein knock-out mice develop vesicle abnormalities [25], thus providing a potential direct link between genetic variability in the gene and a biological pathway involved in the disease. Less is known regarding *NUCKS1*; it has been described to be a nuclear protein, containing casein kinase II and cyclin-dependant kinases phosphorylation sites and to be highly expressed in the cardiac muscle [26]; but an involvement in PD pathogenesis has yet to be suggested.

Table 1. Sample size and genotyping platform for the cohorts included in stage 1 (top set of rows), stage 2 (middle set of rows), and independent replication (bottom row).

Cohort	Controls	Cases	Genotyping platform
United Kingdom	5,200	1,705	Illumina 660W-Quad
USA-NIA	3,034	971	Illumina HumanHap 550
USA-dbGAP	857	876	Illumina 370 K
German	944	742	Illumina HumanHap550
French	1,984	1039	Illumina 610-Quad
Total Stage 1	**12,019**	**5,333**	
Icelandic	1,427	479	Illumina HumanHap 300
Dutch	2,024	772	Illumina 610-Quad
USA	2,215	2,807	ImmunoChip
United Kingdom	1,864	1,271	ImmunoChip
Dutch	402	304	ImmunoChip
French	363	267	ImmunoChip
German	712	1,153	ImmunoChip
Total Stage 2	**9,007**	**7,053**	
Stage 1+Stage 2	**21,026**	**12,386**	
Do et al- USA	**29,624**	**3,426**	

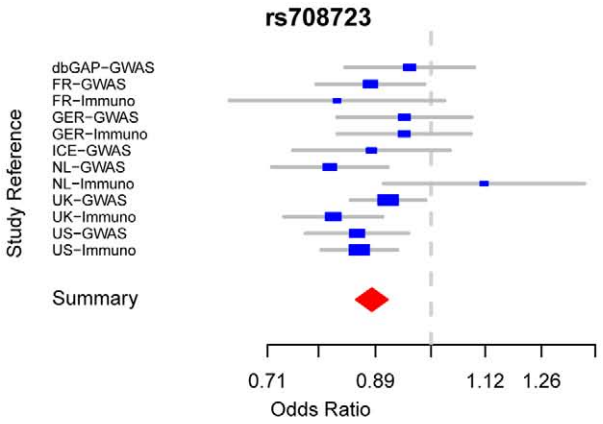
doi:10.1371/journal.pgen.1002142.t001

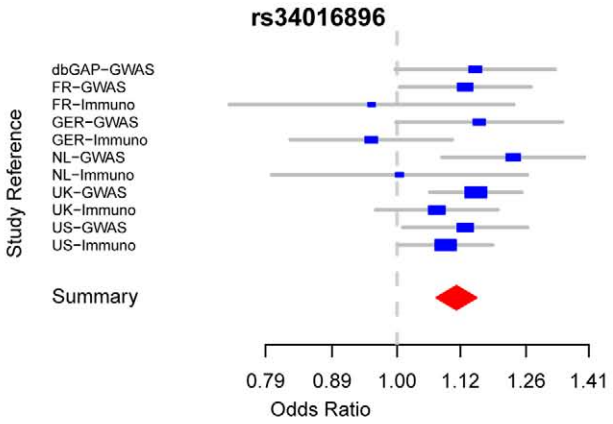
At the 16p11/*STX1B* locus, notwithstanding the fact that other genes are in the associated region, *STX1B* is the most plausible candidate. It has been previously shown to be directly implicated in the process of calcium-dependent synaptic transmission in rat brain [17], having been suggested to play a role in the excitatory pathway of synaptic transmission. Since parkin, encoded by *PARK2*, negatively regulates the number and strength of excitatory synapses [27] , it makes *STX1B* a very interesting candidate from a biologic perspective.

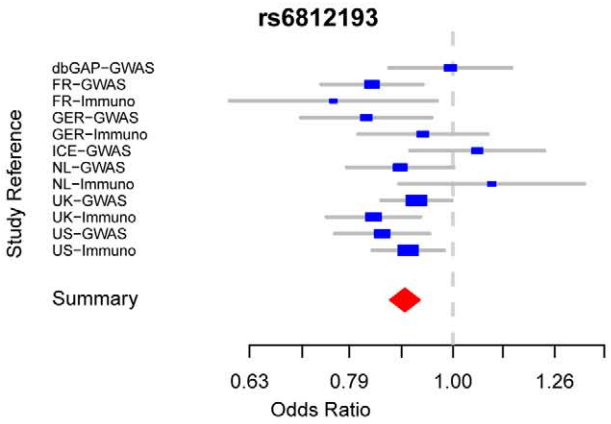
FGF20 at 8p22 has been suggested to be involved in PD [16], albeit negative results in smaller cohorts have followed the original finding [28]. *FGF20* is a neurotrophic factor that exerts strong neurotrophic properties within brain tissue, and regulates central nervous development and function [29]. It is preferentially expressed in the substantia nigra [30], and it has been reported to be involved in dopaminergic neurons survival [30].

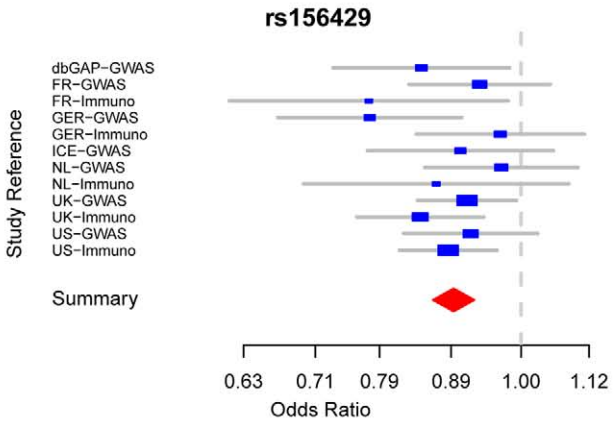
The ImmunoChip data provide limited resolution for the detection of multiple independent association signals in these regions. A previous study [31] reported some evidence of allelic heterogeneity at the 1q32/*PARK16* locus but the ImmunoChip data do not support this result. A previous study [11] also reported two independent associations at the 4q22/*SNCA* locus and our data are consistent with this scenario. However, the newly reported secondary association (rs2245801) is in low LD ($r^2 = 0.21$) with rs2301134, the SNP reported in [11] as an independent association. Taken together, these findings suggest that at least three independent associations exist at *SNCA*/4q22. A more exhaustive fine-mapping analysis using either sequencing of large cohorts or targeted genotyping arrays will also be required to fully explore this locus.

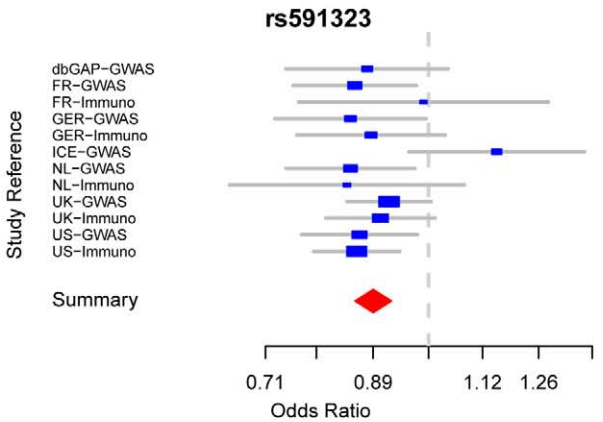
As yet, we do not know which of the variants and which genes within each region are exerting the pathogenic effect. We cannot exclude that some of the currently reported variants are in fact tagging high penetrance, but rare, mutations [32]. Nevertheless, the successful identification of these 16 risk loci further demonstrates the power of the GWA study design, even in the context of disorders like

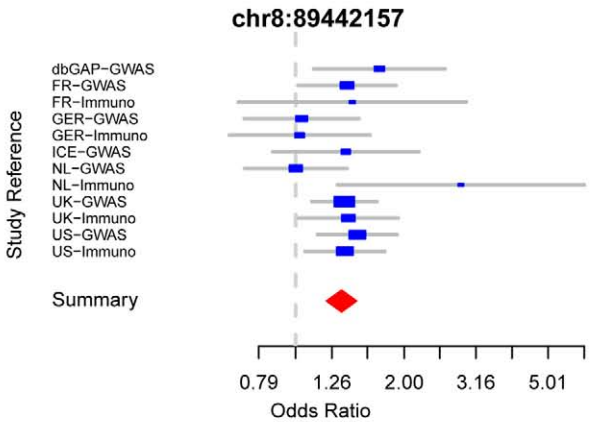

rs708723
dbGAP−GWAS
FR−GWAS
FR−Immuno
GER−GWAS
GER−Immuno
ICE−GWAS
NL−GWAS
NL−Immuno
UK−GWAS
UK−Immuno
US−GWAS
US−Immuno
Summary
Study Reference
0.71
0.89
1.12
1.26
Odds Ratio


rs34016896
dbGAP−GWAS
FR−GWAS
FR−Immuno
GER−GWAS
GER−Immuno
NL−GWAS
NL−Immuno
UK−GWAS
UK−Immuno
US−GWAS
US−Immuno
Summary
Study Reference
0.79
0.89
1.00
1.12
1.26
1.41
Odds Ratio


rs6812193
dbGAP−GWAS
FR−GWAS
FR−Immuno
GER−GWAS
GER−Immuno
ICE−GWAS
NL−GWAS
NL−Immuno
UK−GWAS
UK−Immuno
US−GWAS
US−Immuno
Summary
Study Reference
0.63
0.79
1.00
1.26
Odds Ratio


rs156429
dbGAP−GWAS
FR−GWAS
FR−Immuno
GER−GWAS
GER−Immuno
ICE−GWAS
NL−GWAS
NL−Immuno
UK−GWAS
UK−Immuno
US−GWAS
US−Immuno
Summary
Study Reference
0.63
0.71
0.79
0.89
1.00
1.12
Odds Ratio


rs591323
dbGAP−GWAS
FR−GWAS
FR−Immuno
GER−GWAS
GER−Immuno
ICE−GWAS
NL−GWAS
NL−Immuno
UK−GWAS
UK−Immuno
US−GWAS
US−Immuno
Summary
Study Reference
0.71
0.89
1.12
1.26
Odds Ratio


chr8:89442157
dbGAP−GWAS
FR−GWAS
FR−Immuno
GER−GWAS
GER−Immuno
ICE−GWAS
NL−GWAS
NL−Immuno
UK−GWAS
UK−Immuno
US−GWAS
US−Immuno
Summary
Study Reference
0.79
1.26
2.00
3.16
5.01
Odds Ratio

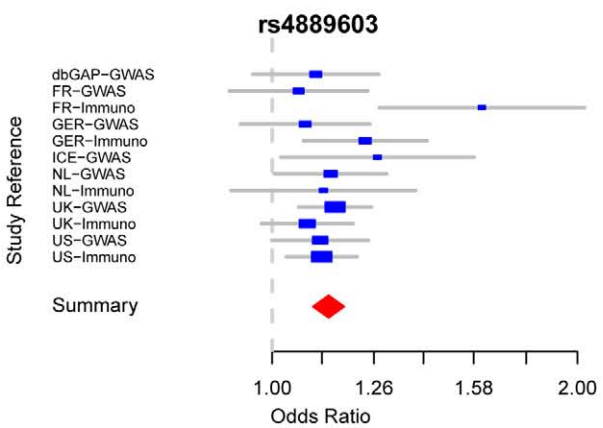

rs4889603
dbGAP−GWAS
FR−GWAS
FR−Immuno
GER−GWAS
GER−Immuno
ICE−GWAS
NL−GWAS
NL−Immuno
UK−GWAS
UK−Immuno
US−GWAS
US−Immuno
Summary
Study Reference
1.00
1.26
1.58
2.00
Odds Ratio

Figure 1. Forest plots detailing effect estimates from the combined analysis of all data contributed by the International Parkinson Disease Genomics Consortium (joint estimates describing constituent effects of Stage 1+Stage 2).
doi:10.1371/journal.pgen.1002142.g001

PD that have a complex genetic component. We therefore expect that further and larger association analyses, perhaps using dedicated high-throughput genotyping arrays like the ImmunoChip, will continue to yield new insights into PD etiology.

Material and Methods

Genotyping and case control cohorts

Participating studies were either genotyped using the ImmunoChip as part of a collaborative agreement with the ImmunoChip Consortium, or as part of previous GWA studies provided by members of the IPDGC or freely available from dbGaP [7,9,10,11]. Genotyping of the UK cases using the Immunochip was undertaken by the WTCCC2 at the Wellcome Trust Sanger Institute which also genotyped the UK control samples. The constituent studies comprising the IPDGC have been described in detail elsewhere [15], although a summary of individual study quality control is available as part of Table S1. In brief all studies followed relatively uniform quality control procedures such as: minimum call rate per sample of 95%, mandatory concordance between self-reported and X-chromosome-heterogeneity estimated sex, exclusion of SNPs with greater than 5% missingness, Hardy Weinberg equilibrium p-values at a minimum of 10^{-7}, minor allele frequencies at a minimum of 1%, exclusion of first degree relatives, and the exclusion of ancestry outliers based on either principal components or multidimensional scaling analyses using either PLINK [33] or EIGENSTRAT [34] to remove non-European ancestry samples. All GWAS studies utilized in this analysis (and in the QTL analyses) were imputed using MACHv1.0.16 [14] to conduct a two-stage imputation based on the August 2009 haplotypes from initial low coverage sequencing of 112 European ancestry samples in the 1000 Genomes Project [35], filtering the data for a minimum imputation quality of (RSQR>0.3) [14]. Logistic regression models were utilized to quantify associations with PD incorporating allele dosages as the primary predictor of disease. Imputed data was analyzed using MACH2DAT, and genotyped SNPs were analyzed using PLINK. All models were adjusted for covariates of components 1 and 2 from either principal components or multidimensional scaling analyses to account for population substructure and stochastic genotypic variation (except in the UK-GWAS data which were not adjusted for population substructure).

Association test statistics

Single SNP test statistics were combined across datasets using a score test methodology, essentially assuming equal odds ratio across cohorts. In addition, fixed and random effects meta-analyses were implemented in R (version 2.11) to confirm that the score test approximation does not affect the interpretation of the results. We also tested the relevant SNPs heterogeneity across cohorts and no significant heterogeneity was detected (Table S2).

Data exchange

We communicated to our colleagues in charge of the independent study (Do et al) the seven SNPs listed in Table 2. For this subset of SNPs they selected the marker with the highest r^2 value on their genotyping platform and provided us with the following summary statistics: odds ratio, direction of effect, standard error for the estimated odds ratio and one degree-of-freedom trend test P-value.

eQTL analysis and methylation analysis

Quantitative trait analyses were conducted to infer effects of risk SNPs on proximal CpG methylation and gene expression. For the five replicated SNP associations (Table 2), all available CpG probes and expression probes within +/−1 MB of the target SNP were

Table 2. Summary statistics for the seven SNPs that pass genome-wide significance ($p<5\times10^{-8}$) in the combined stage 1+2 analysis and that have either not been reported in published PD association studies.

					Stage 1		Stage 2		Stage 1+2	Do et al		Combined
SNP	Chrom	Gene(s)	Alleles	MAF	OR (95%CI)	P	OR (95%CI)	P	P	OR (95%CI)	P	P
rs708723	1q32	*RAB7L1/PARK16*	T>C	0.439	0.905 (0.862–0.95)	6.68×10^{-5}	0.863 (0.824–0.905)	9.47×10^{-10}	1.00×10^{-12}	0.758 (0.65–0.88)	2.12×10^{-6}	8.82×10^{-15}
rs34016896	3q26	*NMD3*	C>T	0.305	1.14 (1.09–1.2)	3.00×10^{-7}	1.08 (1.02–1.14)	0.00399	1.81×10^{-8}	1.002 (0.95–1.06)	0.954	1.31×10^{-6}
rs6812193	4q21	*STBD1*	C>T	0.36	0.886 (0.843–0.932)	2.52×10^{-6}	0.906 (0.864–0.95)	5.29×10^{-5}	7.46×10^{-10}	0.839 (0.79–0.89)	7.55×10^{-10}	1.17×10^{-17}
rs156429	7p15	*GPNMB*	A>G	0.403	0.894 (0.849–0.942)	2.15×10^{-5}	0.893 (0.852–0.937)	3.86×10^{-6}	3.27×10^{-10}	0.901 (0.85–0.95)	0.000193	3.05×10^{-13}
rs591323	8p22	*FGF20*	G>A	0.271	0.884 (0.836–0.935)	1.59×10^{-5}	0.875 (0.83–0.923)	$8.49E\times10^{-7}$	7.45×10^{-11}	0.932 (0.88–0.99)	0.023	1.92×10^{-11}
chr8:89442157	8q21	*MMP16*	C>T	0.0247	1.38 (1.21–1.57)	1.10×10^{-6}	1.29 (1.12–1.49)	0.000451	2.26×10^{-9}	0.969 (0.86–1.09)	0.589	2.36×10^{-5}
rs4889603	16p11	*STX1B*	A>G	0.413	1.12 (1.06–1.18)	4.13×10^{-5}	1.15 (1.1–1.21)	8.21×10^{-9}	2.66×10^{-12}	1.070 (1.01–1.13)	0.014	6.98×10^{-13}

1q32/PARK16 has been reported previously but is included because these data provide for the first time unequivocal evidence of association. P-values are computed using a one-degree-of-freedom regression trend test, including two principal components as covariates and combining the results across cohorts using a score test methodology. P-values are two-tailed and odds ratios are reported for the minor alleles. The notation X>Y indicates that X is the major allele and Y the minor allele. Allele frequencies were estimated using the UK control data. OR: odds ratio.
doi:10.1371/journal.pgen.1002142.t002

Table 3. Significant eQTL associations ($p<0.01$) between the five SNPs with positive replication data (Table 2) and proximal (cis) changes in gene expression/methylation in frontal cortex and cerebellar tissue.

Assay	Region	SNP	Region	Gene Tagged by Probe	Illumina Probe	Alleles	Effect Estimate	Standard Error	Unadjusted P	False Discovery Rate Adjusted P
Expression	Frontal Cortex	rs156429	7p15/*GPNMB*	*NUPL2*	ILMN_1789616	A>G	0.083	0.018	3.6E-06	1.0E-04
		rs156429	7p15/*GPNMB*	*NUPL2*	ILMN_2115154	A>G	0.078	0.017	3.1E-06	1.0E-04
		rs708723	1q32/*PARK16*	*NUCKS1*	ILMN_1680692	T>C	0.155	0.03	1.8E-07	1.5E-05
		rs708723	1q32/*PARK16*	*RAB7L1*	ILMN_1813685	T>C	−0.062	0.018	7.2E-04	1.2E-02
		rs4889603	16p11/*STX1B*	*ZNF668*	ILMN_1739236	A>G	0.062	0.015	4.1E-05	8.7E-04
		rs4889603	16p11/*STX1B*	*MYST1*	ILMN_1804679	A>G	−0.053	0.018	3.4E-03	4.8E-02
	Cerebellum	rs156429	7p15/*GPNMB*	*NUPL2*	ILMN_1789616	A>G	0.133	0.025	1.0E-07	3.7E-06
		rs156429	7p15/*GPNMB*	*NUPL2*	ILMN_2115154	A>G	0.131	0.023	1.2E-08	1.0E-06
		rs708723	1q32/*PARK16*	*NUCKS1*	ILMN_1680692	T>C	0.13	0.029	5.3E-06	1.1E-04
		rs708723	1q32/*PARK16*	*RAB7L1*	ILMN_1813685	T>C	−0.106	0.02	1.3E-07	3.7E-06
		rs4889603	16p11/*STX1B*	*ZNF668*	ILMN_1739236	A>G	0.075	0.02	1.3E-04	2.3E-03
		rs4889603	16p11/*STX1B*	*BCL7C*	ILMN_2371147	A>G	0.066	0.022	2.6E-03	3.8E-02
Methylation	Frontal Cortex	rs156429	7p15/*GPNMB*	*GPNMB*	cg17274742	A>G	−0.027	0.005	5.1E-07	3.2E-05
		rs156429	7p15/*GPNMB*	*GPNMB*	cg22932819	A>G	−0.009	0.002	1.6E-07	1.3E-05
		rs6812193	4q21/*STBD1*	*GENX-3414*	cg17010112	C>T	0.008	0.002	9.4E-04	3.0E-02
		rs708723	1q32/*PARK16*	*FLJ32569*	cg14159672	T>C	−0.219	0.022	3.1E-24	3.9E-22
		rs708723	1q32/*PARK16*	*FLJ32569*	cg14893161	T>C	−0.176	0.017	3.9E-25	9.6E-23
		rs4889603	16p11/*STX1B*	*BCL7C*	cg07896225	A>G	−0.002	0.001	9.7E-04	3.0E-02
		rs4889603	16p11/*STX1B*	*STX1B*	cg25033993	A>G	0.012	0.003	8.2E-05	3.4E-03
	Cerebellum	rs156429	7p15/*GPNMB*	*GPNMB*	cg17274742	A>G	−0.015	0.003	2.1E-06	1.3E-04
		rs708723	1q32/*PARK16*	*FLJ32569*	cg14159672	T>C	−0.246	0.023	3.0E-27	3.7E-25
		rs708723	1q32/*PARK16*	*FLJ32569*	cg14893161	T>C	−0.202	0.018	2.6E-28	6.4E-26

doi:10.1371/journal.pgen.1002142.t003

investigated as candidate QTL associations in frontal cortex and cerebellar tissue samples. 399 samples were assayed for genome-wide gene expression on Illumina HumanHT-12 v3 Expression Beadchips and 292 samples were assayed using Infinium HumanMethylation27 Beadchips, both per manufacturer's protocols in each brain region. A more in depth description of the sample series comprising the QTL analyses, relevant laboratory procedures and quality requirements may be found in [15]. The QTL analysis utilized multivariate linear regression models to estimate effects of allele dosages per SNP on expression and methylation levels adjusted for covariates of age at death, gender, the first 2 component vectors from multi-dimensional scaling, post mortem interval (PMI), brain bank from where the

Table 4. Estimated PD risk profile for the five cohorts genotyped using the Immunochip.

			1st quintile		*2nd quintile*		*3rd quintile*		*4th quintile*		*5th quintile*	
Study	*Trend P-value*	*AUC*	*OR*	*95% CI*	*OR*	*95% CI*	*OR*	*95% CI*	*OR*	*95% CI*	*OR*	*95% CI*
USA	<2E-16	0.614	1	–	1.54	1.29–1.84	1.92	1.61–2.29	2.21	1.85–2.65	3.03	2.52–3.64
UK	<2e-16	0.636	1	–	1.34	1.05–1.71	1.79	1.41–2.28	2.35	1.86–2.99	3.11	2.46–3.96
Germany	1.29E-11	0.692	1	–	1.32	0.98–1.79	1.88	1.38–2.58	1.88	1.38–2.56	2.57	1.88–3.53
France	5.19E-13	0.675	1	–	1.69	0.99–2.92	1.13	0.65–1.98	3.30	1.95–5.67	5.92	3.42–10.52
Netherlands	5.08E-05	0.601	1	–	1.06	0.65–1.74	1.35	0.83–2.20	1.91	1.18–3.11	2.36	1.45–3.86
Combined	<2E-16	0.645	1	–	1.43	1.26–1.61	1.79	1.58–2.02	2.22	1.96–2.50	3.02	2.67–3.42
% Cases per Quintile				**37.90**		**46.06**		**51.15**		**56.56**		**63.75**

Risk scores for the 16 confirmed loci were computed using the odds ratio estimated from the genome-wide case-control genotype data. Individuals were split into quintile on the basis of their risk scores. The odds ratios quantify the effect of the computed risk quintile on the probability of being a PD case (one-degree-of-freedom logistic trend test with the PD status as a binary outcome variable and the quintiles, coded as 1–5, as covariates). The first quantile group was taken as a reference group. OR: odds ratio, CI: confidence interval.

doi:10.1371/journal.pgen.1002142.t004

samples were provided and in which preparation/hybridization batch the samples were processed. A total of 670 candidate QTL associations were tested: 87 expression QTLs in the cerebellum samples, 85 expression QTLs in the frontal cortex samples, 249 methylation QTLs in the cerebellum samples and 249 methylation QTLs in the frontal cortex samples. Multiple test correction was undertaken using false discovery rate adjusted p-values$<$0.05 to dictate significance, with the p-value adjustment undertaken in each series separately, stratified by brain region and assay. A complete list of all QTL associations tested is included in Table S3.

Supporting Information

Table S1 Summary of results for fixed and random effects meta-analysis, as estimates of effect heterogeneity across cohorts and SNP used at the Do et al replication stage.
(XLSX)

Table S2 Summary of the quality control parameters applied to the GWA datasets included in this study.
(XLSX)

Table S3 Complete list of tested QTL associations (expression and methylation).
(XLSX)

Text S1 Membership of the Wellcome Trust Case Control Consortium 2.
(DOC)

Acknowledgments

This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Maryland (http://biowulf.nih.gov). DNA panels and samples from the NINDS Human Genetics Resource Center DNA and Cell Line Repository (http://ccr.coriell.org/ninds) were used in this study, as well as clinical data. The submitters that contributed samples are acknowledged in detailed descriptions of each panel (http://ccr.coriell.org/sections/Collections/NINDS/?SsId = 10).

The authors thank The French Parkinson's Disease Genetics Study Group: Y. Agid, M. Anheim, A.-M. Bonnet, M. Borg, A. Brice, E. Broussolle, J.-C. Corvol, Ph. Damier, A. Destée, A. Dürr, F. Durif, S. Klebe, E. Lohmann, M. Martinez, P. Pollak, O. Rascol, F. Tison, C. Tranchant, M. Vérin, F. Viallet, and M. Vidailhet. The authors thank the members of the French 3C consortium: Drs Annick Alpérovitch, Claudine Berr, Christophe Tzourio, and Jean-Charles Lambert for giving us the possibility to use part of the 3C cohort and Drs. M. Lathrop and D. Zelenika for their support in generating the genome-wide molecular data.

The UK brain samples for the gene expression studies were obtained from the MRC Sudden Death Brain Bank in Edinburgh. This study makes use of GWA data generated by the Wellcome Trust Case-Control consortium 2 (WTCCC2) on UK PD cases and on UK controls from the 1958 Birth Cohort (58BC) and National Blood Service (NBS). UK population control data was made available through WTCCC1. We thank Jeffrey Barrett for assistance with the design of the Immunochip.

The authors of this manuscript are the following: Vincent Plagnol[1], Michael A. Nalls[2], Jose M. Bras[3], Dena G. Hernandez[2,3], Manu Sharma[4], Una-Marie Sheerin[3], Mohamad Saad[4,5], Javier Simón-Sánchez[6], Claudia Schulte[7,8], Suzanne Lesage[9,10,11], Sigurlaug Sveinbjörnsdóttir[12,13,14], Philippe Amouyel[15,16], Sampath Arepalli[1], Gavin Band[17], Roger A. Barker[18], Céline Bellinguez[17], Yoav Ben-Shlomo[19], Henk W. Berendse[20], Daniela Berg[7,8], Kailash Bhatia[21], Rob M. A. [de Bie][22], Alessandro Biffi[23,24,25], Bas Bloem[26], Zoltan Bochdanovits[6], Michael Bonin[27], Kathrin Brockmann[7,8], Janet Brooks[1], David J. Burn[28], Gavin Charlesworth[3], Honglei Chen[29], Patrick F. Chinnery[30], Sean Chong[2], Carl E. Clarke[31,32,33], Mark R. Cookson[2], J. Mark Cooper[34], Jean Christophe Corvol[9,10,11,35], Carl Counsell[36], Philippe Damier[37], Jean-François Dartigues[38], Panos Deloukas[39], Günther Deuschl[40], David T. Dexter[41], Karin D. van Dijk[20], Allissa Dillman[2], Frank Durif[42], Alexandra Dürr[8,9,10,43], Sarah Edkins[39], Jonathan R. Evans[44], Thomas Foltynie[45], Colin Freeman[17], Jianjun Gao[29], Michelle Gardner[3], J. Raphael Gibbs[2,3], Alison Goate[46], Emma Gray[39], Rita Guerreiro[3], Ómar Gústafsson[47], Clare Harris[36], Garrett Hellenthal[17], Jacobus J. van Hilten[48], Albert Hofman[49], Albert Hollenbeck[50], Janice Holton[51], Michele Hu[52], Xuemei Huang[53], Heiko Huber[7,8], Gavin Hudson[30], Sarah E. Hunt[39], Johanna Huttenlocher[15], Thomas Illig[54], Pálmi V. Jónsson[55], Cordelia Langford[44], Andrew Lees[51], Peter Lichtner[56], Patricia Limousin[57], Grisel Lopez[58], Delia Lorenz[40], Alisdair McNeill[34], Catriona Moorby[31], Matthew Moore[2], Huw Morris[59], Karen E. Morrison[31,60], Ese Mudanohwo[61], Sean S. O'Sullivan[51], Justin Pearson[59], Richard Pearson[17], Joel S. Perlmutter[46], Hjörvar Pétursson[27,47], Matti Pirinen[17], Pierre Pollak[62], Bart Post[26], Simon Potter[39], Bernard Ravina[63], Tamas Revesz[51], Olaf Riess[27], Fernando Rivadeneira[49,64], Patrizia Rizzu[6], Mina Ryten[3], Stephen Sawcer[65], Anthony Schapira[34], Hans Scheffer[66], Karen Shaw[51], Ira Shoulson[67], Ellen Sidransky[58], Rohan de Silva[3], Colin Smith[68], Chris C. A. Spencer[69], Hreinn Stefánsson[47], Stacy Steinberg[47], Joanna D. Stockton[31], Amy Strange[17], Zhan Su[17], Kevin Talbot[69], Carlie M. Tanner[70], Avazeh Tashakkori-Ghanbaria[39], François Tison[71], Daniah Trabzuni[3], Bryan J. Traynor[2], André G. Uitterlinden[49,64], Jana Vandrovcova[3], Daan Velseboer[22], Marie Vidailhet[9,10,11], Damjan Vukcevic[17], Robert Walker[68], Bart van de Warrenburg[26], Michael E. Weale[72], Mirdhu Wickremaratchi[73], Nigel Williams[59], Caroline H. Williams-Gray[18], Sophie Winder-Rhodes[74], Kári Stefánsson[47], Maria Martinez[4,5], Peter Donnelly[17], Andrew B. Singleton[2], John Hardy[3], Peter Heutink[6], Alexis Brice[9,10,11,43], Thomas Gasser[7,8], Nicholas W. Wood[1,3]*

1 UCL Genetics Institute, University College London, London, United Kingdom, **2** Laboratory of Neurogenetics. National Institute on Aging, National Institutes of Health, Bethesda, Maryland, United States of America, **3** Department of Molecular Neuroscience, Institute of Neurology, University College London, London, United Kingdom, **4** Institut National de la Sante et de la Recherche Medicale, UMR 1043, Centre de Physiopathologie de Toulouse-Purpan, Toulouse, France, **5** Paul Sabatier University, Toulouse, France, **6** Department of Clinical Genetics, Section of Medical Genomics, VU University Medical Centre, Amsterdam, The Netherlands, **7** Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, University of Tübingen, Tübingen, Germany, **8** Deutsches Zentrum fur Neurodegenerative Erkrangungen (German Center for Neurodegenerative Diseases), Tübingen, Germany, **9** Institut National de la Sante et de la Recherche Medicale, UMR_S975 (Formerly UMR_S679), Paris, France, **10** Université Pierre et Marie Curie-Paris, Centre de Recherche de l'Institut du Cerveau et de la Moelle épinière, UMR-S975, Paris, France, **11** Centre National de la Recherche Scientifique, UMR 7225, Paris, France, **12** Department of Neurology, Landspítali University Hospital, Reykjavík, Iceland, **13** Department of Neurology, Mid Essex Hospital, Broomfield Hospital, Chelmsford, Essex, United Kingdom, **14** Queen Mary College, University of London, London, United Kingdom, **15** Institut National de la Sante et de la Recherche Medicale, U744, Lille, France, **16** Institut Pasteur de Lille, Université de Lille Nord, Lille, France, **17** Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, United Kingdom, **18** Department of Neurology, Addenbrooke's Hospital, University of Cambridge, Cambridge, United Kingdom, **19** Department of Social Medicine, Bristol University, Bristol, United Kingdom, **20** Department of Neurology and Alzheimer Center, VU University Medical Center, Amsterdam, The Netherlands, **21** Department of Motor Neuroscience, University College London Institute of Neurology, London, United Kingdom, **22** Department of Neurology, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands, **23** Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, United States of America, **24** Department of Neurology, Massachusetts General Hospital, Boston, Massachusetts, United States of America, **25** Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, United States of America, **26** Department of Neurology, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands, **27** Department of Medical Genetics, Institute of Human Genetics, University of Tübingen, Tübingen, Germany, **28** Newcastle University Clinical Ageing Research Unit, Campus for Ageing and Vitality, Newcastle upon Tyne, United Kingdom, **29** Epidemiology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, North Carolina, United States of America, **30** Neurology Department, The Medical School, Newcastle upon Tyne, Newcastle University, United Kingdom, **31** School of Clinical and Experimental Medicine, University of Birmingham, Edgbaston, Birmingham, United Kingdom, **32** Department of Neurology, City Hospital, Sandwell, United Kingdom, **33** West

Birmingham Hospitals NHS Trust, Birmingham, United Kingdom, **34** Department of Clinical Neurosciences, University College London Institute of Neurology, London, United Kingdom, **35** Institut National de la Sante et de la Recherche Medicale, CIC-9503, Hôpital Pitié-Salpêtrière, Paris, France, **36** University of Aberdeen, Division of Applied Health Sciences, Population Health Section, Aberdeen, United Kingdom, **37** Centre Hospitalier Universitaire Nantes, CIC0004, Service de Neurologie, Nantes, France, **38** Institut National de la Sante et de la Recherche Medicale, U897, Université Victor Segalen, Bordeaux, France, **39** Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, **40** Klinik für Neurologie, Universitätsklinikum Schleswig-Holstein, Campus Kiel, Christian-Albrechts-Universität Kiel, Kiel, Germany, **41** Parkinson's Disease Research Group, Faculty of Medicine, Imperial College London, London, United Kingdom, **42** Service de Neurologie, Hôpital Gabriel Montpied, Clermont-Ferrand, France, **43** AP-HP, Pitié-Salpêtrière Hospital, Department of Genetics and Cytogenetics, Paris, France, **44** Cambridge Centre for Brain Repair, University of Cambridge, Cambridge, United Kingdom, **45** Institute of Neurology, University College London, London, United Kingdom, **46** Department of Psychiatry, Department of Neurology, Washington University School of Medicine, St. Louis, Missouri, United States of America, **47** 14 deCODE genetics, Reykjavik, Iceland, **48** Department of Neurology, Leiden University Medical Center, Leiden, The Netherlands, **49** Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands, **50** American Association of Retired Persons, Washington DC, United States of America, **51** Queen Square Brain Bank for Neurological Disorders, Institute of Neurology, University College London, London, United Kingdom, **52** Department of Clinical Neurology, John Radcliffe Hospital, Oxford, United Kingdom, **53** Departments of Neurology, Radiology, Neurosurgery, Pharmacology, Kinesiology, and Bioengineering, Pennsylvania State University–Milton S. Hershey Medical Center, Hershey, Pennsylvania, United States of America, **54** Institute of Epidemiology, Helmholtz Zentrum München, German Research Centre for Environmental Health, Neuherberg, Germany, **55** Department of Geriatrics, Landspítali University Hospital, Reykjavík, Iceland, **56** Institute of Human Genetics, Helmholtz Zentrum München, German Research Centre for Environmental Health, Neuherberg, Germany, **57** Sobell Department, Unit of Functional Neurosurgery, University College London Institute of Neurology, London, United Kingdom, **58** Section on Molecular Neurogenetics, Medical Genetics Branch, NHGRI, National Institutes of Health, Bethesda, Maryland, United States of America, **59** Medical Research Council Centre for Neuropsychiatric Genetics and Genomics, Cardiff University School of Medicine, Cardiff, United Kingdom, **60** Neurosciences Department, Queen Elizabeth Hospital, University Hospitals Birmingham NHS Foundation Trust, Birmingham, United Kingdom, **61** Neurogenetics Unit, University College London, Institute of Neurology/National Hospital for Neurology and Neurosurgery, London, United Kingdom, **62** Service de Neurologie, Centre Hospitalier Universitaire de Grenoble, Grenoble, France, **63** Translational Neurology, Biogen Idec, Cambridge, Massachusetts, United States of America, **64** Department of Internal Medicine, Erasmus Medical Center, Rotterdam, The Netherlands, **65** University of Cambridge, Department of Clinical Neurosciences, Addenbrooke's Hospital, Cambridge, United Kingdom, **66** Department of Human Genetics, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands, **67** Department of Neurology, University of Rochester, Rochester, New York, United States of America, **68** Department of Pathology, University of Edinburgh, Edinburgh, United Kingdom, **69** University of Oxford, Department of Clinical Neurology, John Radcliffe Hospital, Oxford, United Kingdom, **70** Clinical Research Department, The Parkinson's Institute and Clinical Center, Sunnyvale, California, United States of America, **71** Service de Neurologie, Hôpital Haut-Lévêque, Pessac, France, **72** Department of Medical and Molecular Genetics, King's College London, London, United Kingdom, **73** Department of Neurology, Cardiff University, Cardiff, United Kingdom, **74** Department of Psychiatry and Medical Research Council/Wellcome Trust Behavioural and Clinical Neurosciences Institute, University of Cambridge, Cambridge, United Kingdom.

Author Contributions

Conceived and designed the experiments: V Plagnol, MA Nalls, M Martinez, P Donnelly, J Hardy, P Heutink, A Brice, T Gasser, AB Singleton, NW Wood. Analyzed the data: V Plagnol, MA Nalls, JM Bras. Contributed reagents/materials/analysis tools: DG Hernandez, M Sharma, U-M Sheerin, J Simón-Sánchez, C Schulte, S Lesage, S Sveinbjörnsdóttir, P Amouyel, S Arepalli, G Band, RA Barker, C Bellinguez, Y Ben-Shlomo, HW Berendse, D Berg, K Brockmann, RMA de Bie, A Brice, A Biffi, B Bloem, Z Bochdanovits, M Bonin, K Bhatia, J Brooks, DJ Burn, G Charlesworth, H Chen, PF Chinnery, S Chong, CE Clarke, A Dürr, A Dillman, DT Dexter, F Tison, F Durif, KD van Dijk, M Saad, MR Cookson, JM Cooper, JC Corvol, C Counsell, P Damier, J-F Dartigues, P Deloukas, G Deuschl, S Edkins, JR Evans, T Foltynie, C Freeman, J Gao, M Gardner, JR Gibbs, R Guerreiro, A Goate, E Gray, O Gústafsson, C Harris, G Hellenthal, JJ van Hilten, A Hofman, A Hollenbeck, J Holten, J Huttenlocher, M Hu, X Huang, H Huber, G Hudson, SE Hunt, T Illig, PV Jónsson, C Langford, A Lees, P Lichtner, P Limimousin, G Lopez, D Lorenz, A McNeill, C Moorby, H Morris, KE Morrison, E Mudhanohwo, SS O'Sullivan, J Pearson, R Pearson, JS Perlmutter, H Pétursson, M Pirinen, P Pollak, B Post, S Potter, B Ravina, T Revesz, O Riess, F Rivadeneira, P Rizzu, M Ryten, S Sawcer, A Schapira, H Scheffer, K Shaw, I Shoulson, E Sidransky, R de Silva, C Smith, CCA Spencer, H Stefánsson, S Steinberg, JD Stockton, A Strange, Z Su, K Talbot, CM Tanner, A Tashakkori-Ghanbaria, D Tison, BJ Traynor, AG Uitterlinden, J Vandrovcova, D Velseboer, M Vidailhet, D Vukcevic, R Walker, B van de Warrenburg, ME Weale, M Wickremaratchi, N Williams, CH Williams-Gray, S Winder-Rhodes, K Stefánsson, M Moore, P Donnelly, AB Singleton, J Hardy, P Heutink, T Gasser, NW Wood. Wrote the manuscript: V Plagnol, MA Nalls, AB Singleton, NW Wood.

References

1. Zimprich A, Müller-Myhsok B, Farrer M, Leitner P, Sharma M, et al. (2004) The PARK8 locus in autosomal dominant parkinsonism: confirmation of linkage and further delineation of the disease-containing interval. American journal of human genetics 74: 11–19.
2. Paisán-Ruíz C, Jain S, Evans W, Gilks W, Simón J, et al. (2004) Cloning of the gene containing mutations that cause PARK8-linked Parkinson's disease. Neuron 44: 595–600.
3. Valente EM, Abou-Sleiman P, Caputo V, Muqit M, Harvey K, et al. (2004) Hereditary early-onset Parkinson's disease caused by mutations in PINK1. Science 304: 1158–1160.
4. Polymeropoulos MH, Lavedan C, Leroy E, Ide SE, Dehejia A, et al. (1997) Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. Science 276: 2045–2047.
5. Kitada T, Asakawa S, Hattori N, Matsumine H, Yamamura Y, et al. (1998) Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. Nature 392: 605–608.
6. Bonifati V, Rizzu P, van Baren M, Schaap O, Breedveld G, et al. (2003) Mutations in the DJ-1 gene associated with autosomal recessive early-onset parkinsonism. Science 299: 256–259.
7. Simón-Sánchez J, Schulte C, Bras J, Sharma M, Gibbs R, et al. (2009) Genome-wide association study reveals genetic risk underlying Parkinson's disease. Nature Genetics 41: 1308–1312.
8. Satake W, Nakabayashi Y, Mizuta I, Hirota Y, Ito C, et al. (2009) Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. Nature Genetics 41: 1303–1307.
9. Saad M, Lesage S, Saint-Pierre A, Corvol J-C, Zelenika D, et al. (2011) Genome-wide association study confirms BST1 and suggests a locus on 12q24 as risk loci for Parkinson's disease in the European population. Human Molecular Genetics 20: 615–627.
10. Simon-Sanchez J, van Hilten J, van de Warrenburg B, Post B, Berendse H, et al. (2011) Genome-wide association study confirms extant PD risk loci among the Dutch. European Journal of Human Genetics aop.
11. (2011) Dissection of the genetics of Parkinson's disease identifies an additional association 5′ of SNCA and multiple associated haplotypes at 17q21. Human Molecular Genetics 20: 345–353.
12. Pankratz N, Wilk J, Latourelle J, DeStefano A, Halter C, et al. (2009) Genomewide association study for susceptibility genes contributing to familial Parkinson disease. Human genetics 124: 593–605.
13. Hamza T, Zabetian C, Tenesa A, Laederach A, Montimurro J, et al. (2010) Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. Nature Genetics 42: 781–785.
14. Li Y, Willer C, Ding J, Scheet P, Abecasis G (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic epidemiology 34: 816–834.

15. Nalls M, Plagnol V, Hernandez D, Sharma M, Sheerin U-M, et al. (2011) Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet 377: 641–649.
16. van der Walt J, Noureddine M, Kittappa R, Hauser M, Scott W, et al. (2004) Fibroblast growth factor 20 polymorphisms and haplotypes strongly influence risk of Parkinson disease. American journal of human genetics 74: 1121–1127.
17. Smirnova T, Stinnakre J, Mallet J (1993) Characterization of a presynaptic glutamate receptor. Science 262: 430–433.
18. Wang G, van der Walt J, Mayhew G, Li Y-J, Züchner S, et al. (2008) Variation in the miRNA-433 binding site of FGF20 confers risk for Parkinson disease by overexpression of alpha-synuclein. American journal of human genetics 82: 283–289.
19. Wider C, Dachsel J, Soto A, Heckman M, Diehl N, et al. (2009) FGF20 and Parkinson's disease: no evidence of association or pathogenicity via alpha-synuclein expression. Movement disorders 24: 455–459.
20. Gilks W, Abou-Sleiman P, Gandhi S, Jain S, Singleton A, et al. (2005) A common LRRK2 mutation in idiopathic Parkinson's disease. Lancet 365: 415–416.
21. Zollner S, Pritchard J (2007) Overcoming the winner's curse: estimating penetrance parameters from case-control data. American journal of human genetics 80: 605–615.
22. Plagnol V, Smyth D, Todd J, Clayton D (2008) Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. Biostatistics (Oxford, England).
23. Shimizu F, Katagiri T, Suzuki M, Watanabe TK, Okuno S, et al. (1997) Cloning and chromosome assignment to 1q32 of a human cDNA (RAB7L1) encoding a small GTP-binding protein, a member of the RAS superfamily. Cytogenetics and cell genetics 77: 261–263.
24. Koshimura K, Ohue T, Akiyama Y, Itoh A, Miwa S (1992) L-dopa administration enhances exocytotic dopamine release in vivo in the rat striatum. Life sciences 51: 747–755.
25. Cabin D, Shimazu K, Murphy D, Cole N, Gottschalk W, et al. (2002) Synaptic vesicle depletion correlates with attenuated synaptic responses to prolonged repetitive stimulation in mice lacking alpha-synuclein. The Journal of neuroscience 22: 8797–8807.
26. Grundt K, Haga IV, Aleporou-Marinou V, Drosos Y, Wanvik B, et al. (2004) Characterisation of the NUCKS gene on human chromosome 1q32.1 and the presence of a homologous gene in different species. Biochemical and biophysical research communications 323: 796–801.
27. Helton T, Otsuka T, Lee M-C, Mu Y, Ehlers M (2008) Pruning and loss of excitatory synapses by the parkin ubiquitin ligase. Proceedings of the National Academy of Sciences of the United States of America 105: 19492–19497.
28. Clarimon J, Xiromerisiou G, Eerola J, Gourbali V, Hellström O, et al. (2005) Lack of evidence for a genetic association between FGF20 and Parkinson's disease in Finnish and Greek patients. BMC neurology 5.
29. Jeffers M, Shimkets R, Prayaga S, Boldog F, Yang M, et al. (2001) Identification of a novel human fibroblast growth factor and characterization of its role in oncogenesis. Cancer research 61: 3131–3138.
30. Ohmachi S, Mikami T, Konishi M, Miyake A, Itoh N (2003) Preferential neurotrophic activity of fibroblast growth factor-20 for dopaminergic neurons through fibroblast growth factor receptor-1c. J Neurosci Res 72: 436–443.
31. Tucci A, Nalls M, Houlden H, Revesz T, Singleton A, et al. (2010) Genetic variability at the PARK16 locus. European journal of human genetics : EJHG 18: 1356–1359.
32. Dickson S, Wang K, Krantz I, Hakonarson H, Goldstein D (2010) Rare Variants Create Synthetic Genome-Wide Associations. PLoS Biol 8: e1000294. doi:10.1371/journal.pbio.1000294.
33. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 81: 559–575.
34. Price A, Patterson N, Plenge R, Weinblatt M, Shadick N, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38: 904–909.
35. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.